Machine-learning derived algorithms for prediction of radiographic progression in early axial spondyloarthritis

R. Garofoli[1], M. Resche-Rigon[2,3], C. Roux[1,4], D. van der Heijde[5], M. Dougados[1,4], A. Moltó[1,4]

[1]ECAMO team, [2]ECSTRRA Team; INSERM U1153: Clinical Epidemiology and Biostatistics, Université de Paris, Paris, France; [3]SBIM, Saint-Louis Hospital, APHP, Paris, France; [4]Rheumatology Department, Cochin Hospital, APHP.5, Paris, France; [5]Leiden University Medical Center (LUMC), Leiden, The Netherlands.

Abstract

Objective

To compare machine learning (ML) to traditional models to predict radiographic progression in patients with early axial spondyloarthritis (axSpA).

Methods

We carried out a prospective French multicentric DESIR cohort study with 5 years of follow-up that included patients with chronic back pain for <3 years, suggestive of axSpA. Radiographic progression was defined as progression at the spine (increase of at least 1 point of mSASSS scores/2 years) or at the sacroiliac joint (worsening of at least one grade of the mNY score between 2 visits). Statistical analyses were based on patients without any missing data regarding the outcome and variables of interest (295 patients).

Traditional modelling: we performed a multivariate logistic regression model (M1); then variable selection with stepwise selection based on Akaike Information Criterion (stepAIC) method (M2), and Least Absolute Shrinkage and Selection Operator (LASSO) method (M3).

ML modelling: using "SuperLearner" package on R, we modelled radiographic progression with stepAIC, LASSO, random forest, Discrete Bayesian Additive Regression Trees Samplers (DBARTS), Generalized Additive Models (GAM), multivariate adaptive polynomial spline regression (polymars), Recursive Partitioning And Regression Trees (RPART) and Super Learner. Accuracy of these models was compared based on their 10-fold cross-validated AUC (cv-AUC).

Results

10-fold cv-AUC for traditional models were 0.79 and 0.78 for M2 and M3, respectively. The three best models in the ML algorithms were the GAM, the DBARTS and the Super Learner models, with 10-fold cv-AUC of: 0.77, 0.76 and 0.74, respectively.

Conclusion

Two traditional models predicted radiographic progression as good as the eight ML models tested in this population.

Key words

machine learning, spondyloarthritis, cohort

Romain Garofoli, MD, MSc
Matthieu Resche-Rigon, MD, PhD
Christian Roux, MD, PhD
Désirée van der Heijde, MD, PhD
Maxime Dougados, MD, PhD
Anna Moltó, MD, PhD

Please address correspondence to:
Romain Garofoli
Rheumatology Department,
Cochin Hospital,
27 rue du Faubourg Saint Jacques,
75014 Paris, France.
E-mail: romaingarofoli@gmail.com

Received on April 15, 2022; accepted in revised form on June 27, 2022.

Competing interests: R. Garofoli received a grant from the Société Française de Rhumatologie to conduct this study.
C. Roux has received research grants and/or honoraria from Alexion, Amgen, Regeneron and UCB;
D. van der Heijde has received consultancy fees from AbbVie, Bayer, Bristol-Myers Sqibb, Cyxone, Eisai, Galapagos, Gilead, Glaxo-Smith-Kline. Janssen, Lilly, Novartis, Pfizer and UCB Pharma, and is director of Imaging Rheumatolgy B.V.
A. Moltó has received honoraria from AbbVie, Biogen, Bristol-Myers Sqibb, Gilead, Janssen, Lilly, MSD, Novartis, Pfizer, Sanofi and UCB; and research grants from Pfizer (Passerelle) and UCB.
M. Resche-Rigon and M. Dougados have declared no competing interests.

Introduction

Axial spondyloarthritis (axSpA) is a chronic multifaceted rheumatic disease that encompasses various clinical presentations, including chronic back pain (mostly inflammatory), peripheral manifestations such as arthritis, enthesitis or dactylitis, and extra-articular manifestations such as psoriasis, uveitis or inflammatory bowel disease (1).

Radiographic progression can also be highly variable between patients and can occur early in the disease or after decades (2-4). Several factors have been classically found to be associated with a higher radiographic progression rate, in particular the appearance of syndesmophytes: smoking, HLAB27, male gender, young age at diagnosis, increased C-reactive protein (CRP), higher disease activity, a physically demanding job ('blue-collar' job), sacroiliac joint (SIJ) MRI inflammation, spinal MRI inflammation and structural lesions at baseline (syndesmophytes or radiographic sacroiliitis) (5-7).

Predictive factors for radiographic progression in axSpA have been identified through use of traditional statistical models, such as logistic regression (3-7). However, these models present some limitations, like the need to include the relevant independent variables (in order to have a performing predictive model) but at the same time the limitation of needing a good number of observations to achieve stable, meaningful results: it is generally accepted that logistic regression needs at least 10 cases per independent variable in the analysis (8). Finally, having too many parameters compared to observations may lead to overfitting (*i.e.* when a model is excessively complex, and reflects too much the sample; such model will overreact to minor fluctuations in the sample data, leading to a poor predictive performance in other data sets) (9).

In order to overcome these limitations and to improve the predictive performance, machine learning (ML) methods have been developed. ML is a subfield of artificial intelligence (AI), and combines computer science and mathematics to develop methods which are able to "learn" from experience (data) and create predictive and prognostic models with high accuracy, reliability, and efficiency (10). ML can model complex relationships between large explanatory features and desired outputs.

These new analytical tools have recently been used in other medical disciplines, such as the field of oncology, neurosurgery and neuro-imaging (11-12). But to date only few studies have applied these models in rheumatology and to our best knowledge none aiming to predict radiographic progression in spondyloarthritis (13-14).

The goal of this study was to compare the accuracy of ML algorithms to traditional models to predict radiographic progression in patients with early axSpA.

Materials and methods

Patients

The Devenir des Spondylarthropathies Indifférenciées Récentes (DESIR) cohort. DESIR (www.lacohortedesir.fr/desir-in-english), NCT01648907, is a prospective longitudinal cohort involving 25 rheumatology centres in France. Participants at the study gave their written informed consent. This study fulfilled the current Good Clinical Practices and has obtained the approval of the appropriate ethical committee. DESIR's characteristics have been described elsewhere (15), but briefly, 708 consecutive adult patients (inclusion period 2007 to April 2010), aged <50 years with chronic but early (>3 months but <3 years) inflammatory back pain (IBP) highly suggestive of SpA according to the rheumatologists' assessment (score ≥5 on a Numerical Rating Scale (NRS) of 0–10 where 0 = not suggestive and 10 = very suggestive of SpA) were included. Visits were scheduled every 6 months for the first 2 years and yearly thereafter. A 15-year follow-up is currently ongoing, but the present analysis focuses on the first 5 years of follow-up. Among these patients, we analysed the data of patients with no missing data regarding the outcome and the variable of interest (see below).

Patient and public involvement

Patients were not involved in the design of this study.

Data collected

Variables collected at each DESIR cohort visit have already been described elsewhere (15): patients' characteristics (age, sex, socio-demographic features, smoking status, employment), SpA clinical features (date of disease onset, peripheral involvement, enthesitis), disease activity (BASDAI, ASDAS and CRP (mg/L)) and severity (BASFI), and local reading imaging (radiographic sacroiliitis, MRI sacroiliitis) were collected at baseline and at each DESIR visit according to the study protocol (detailed protocol and collected variables information available online at www.lacohortedesir.fr/desir-in-english). In addition, all images for baseline, 2 and 5 years were centrally read by 3 readers per modality (for details see below), who were blinded for time of acquisition of the images and clinical information. For each imaging modality, scores from readers were combined: for continuous outcomes the mean of the available readers was calculated; for binary outcomes the score agreed by 2 out of the 3 readers was retained. The database used for the analyses was locked in June 2018.

Radiographic scores

- mSASSS

Radiographic damage of the spine was assessed by the mSASSS score (16), that ranges from 0 to 72. It has previously been shown that this score was useful and reliable for assessing radiographic damage in spondyloarthritis and for detecting changes over time (17).

- mNY

Radiographic damage of the SIJ was assessed by the modified New York (mNY) scoring system. According to this scoring method, the reader must score each sacroiliac joint from 0 to 4 (0 = no disease, 1 = suspicious for sacroiliitis, 2 = small localised areas with erosions or sclerosis without alteration in joint width, 3 = moderate/advanced sacroiliitis with one or more of erosions, evidence of sclerosis, widening, narrowing or partial ankylosis; 4 = total ankylosis (18). The final score (per SIJ) ranges from 0 to 4. A total SIJ radiographic score per individual can be calculated, which ranges from 0 to 8. The mNY criteria (binary criteria) are fulfilled at the individual level (i.e. for 2 SIJ) if there is at least a grade 2 bilaterally or a unilateral grade 3 or 4 (18).

Definitions

- Radiographic progression

Spine: radiographic progression at the spine was defined as the increase of at least 1 mSASSS unit per 2 years as it is clinically relevant (3-4, 19).

Sacroiliac joints: radiographic progression of the SIJ was defined as the increase of at least one grade of the mNY score (continuous variable) between 2 visits (visits at baseline, 2y and 5y), according to the mNY scoring method, except from an increase from 0 to 1, which was not considered significant (5).

Patients were defined as "progressors" if presenting at 5 years with radiographic progression either at the SIJ or the spine.

Statistical analysis

All analyses were performed on R, v. 3.6.0. First, we performed a bivariate analysis between radiographic progression and variables collected at baseline that have been classically reported to be relevant for radiographic progression in the literature (See Supplementary Table S1). For quantitative variables, Student t-test and Mann-Whitney/Wilcoxon test were used as appropriate (20-21). For binary variables, Chi-Square and Fisher's test were used as appropriate (22). For non-binary categorical variables, Fisher's test was used.

- Traditional models

We performed a multiple logistic regression to predict radiographic progression (M1 model). All variables with a $p<0.4$ in the bivariate analysis were included in the model (23). Then, two different methods to select the variables were applied:

- stepwise selection based on the AIC (step AIC (AIC: Akaike Information Criterion)) method, backward and forward: basically, step AIC respectively removes or add variables in the model until getting the model with the best (the lowest) AIC in a backward or forward way (24): M2 model;
- LASSO (Least Absolute Shrinkage and Selection Operator) method: LASSO method penalises the likelihood of the model in order to force certain regression coefficients to be set to zero. It has been developed notably to handle overfitting (25): M3 model.

For each model, we calculated the 10-fold cross-validated Area Under the Curve (10-fold cv-AUC) (26), using the R packages "pROC", "cvAUC" and "caret" (27).

ML approach

ML algorithms are usually characterised according to different paremeters: parametric *versus* non-parametric, supervised *versus* unsupervised, unique *versus* ensemble algorithms. The Super Learner (SL) is a supervised ensemble of ML algorithms using simultaneously parametric and non-parametric methods (28), and can be used for selecting the optimal prediction algorithm among a set of candidate algorithms via k-fold cross-validation (29). Moreover, it can further capitalise on the performance of all candidate algorithms included in its library building an aggregate algorithm defined as an optimal weighted combination of all candidate algorithms. In practice, candidate algorithms in the SL library were trained and ranked according to their average estimated risk and the algorithm with the least average estimated risk was identified. We defined the following library of ML models: step AIC, LASSO, Random forest (an ensemble learning method for classification, among the most used ML methods; they operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees (30-32)), Discrete Bayesian Additive Regression Trees Samplers (DBARTS, which creates a sampler object fitting a Bayesian Additive Regression Tree (BART) model, which is a Bayesian "sum-of-trees" model in which each tree is constrained by a prior to be a weak learner; it is a nonparametric regression approach defined by a statistical model (a priori and a likelihood) that uses dimensionally adaptive random basis elements, and

it can be used for model-free variable selection (33)), Generalized Additive Models (GAM, statistical models in which the usual linear relationship between the response and predictors are replaced by several non-linear smooth functions to model and capture the non-linearities in the data (34-35)), multivariate adaptive polynomial spline regression (polymars, a non-parametric regression analysis technique (like an extension of linear models) that automatically models nonlinearities and interactions between variables (36-37)), Recursive Partitioning And Regression Trees (RPART, which can build regression models of a very general structure using a two-stage procedure: first the single variable which best splits the data into two groups is found. The data is separated, and then this process is applied separately to each sub-group, and so on recursively until the subgroups either reach a minimum size or until no improvement can be made. The second stage of the procedure consists of using cross-validation to trim back the full tree. Finally, the resulting models can be represented as binary trees (12, 30-31).

A comparison of 10-fold cross-validated AUC of different models at the same time is performed by SL (31). Finally, the SL is a unique combination of the different ML models, automatically computed by the statistical package ("SuperLearner" R package) (28-31).

Finally, traditional and ML models were compared by their 10-fold cross-validated AUC, which is currently one of the most popular discrimination indexes (38-39). We also calculated for each model: the Brier score and the Hosmer-Lemeshow goodness of fit (40–42).

Handling of missing data

All these models were performed only on the population without any missing data on the outcome or on the variables of interest.

We performed a sensitivity analysis based on all patients after multiple imputation of all missing data by chained equations using "MICE" package on R which takes into account all information available for each patient to impute missing data on the other variables (43). We imputed 60 data sets with 500 iterations and made a sensitivity analysis with a manual backward method (M4 model).

Table I. Baseline characteristics of patients included in the analysis (*e.g.* with no missing data on the outcome).

	All included patients n=295	Non progressors n=207	Progressors n=88	*p*-value
General characteristics				
Age, mean (SD)	34.5 (6.8)	34.1	35.3	0.272
Women, n (%)	148 (50.2%)	112 (54.1%)	36 (40.1%)	0.053
BMI, mean (SD)	24.2 (3.3)	24.1	24.5	0.708
Current smoker, n (%)	110 (37.3%)	66 (31.9%)	43 (48.9%)	<0.001
Profession: blue-collar job, n (%)	30 (10.2%)	17 (8.2%)	13 (14.8%)	0.112
Past history information				
HLA B27 positive, n (%)	188 (63.7%)	133 (64.3%)	55 (62.5%)	0.521
Past history of enthesitis, n (%)	173 (58.6%)	128 (61.8%)	45 (51.1%)	0.053
Past history of dactylitis, n (%)	40 (13.6%)	31 (15.0%)	9 (10.2%)	0.141
Past history of arthritis, n (%)	82 (27.8)	59 (28.5%)	23 (26.1%)	0.508
Uveitis, n (%)	24 (8.1%)	15 (7.2%)	9 (10.2%)	0.419
Inflammatory bowel disease, n (%)	15 (5.1%)	11 (5.3%)	4 (4.5%)	0.826
Psoriasis, n (%)	53 (18.0%)	38 (18.4%)	15 (17.0%)	0.977
Disease activity and function				
BASDAI score, mean (SD)	44.0 (18.8)	44.3	43.2	0.467
ASDAS-CRP score, mean (SD)	2.5 (0.8)	2.4	2.7	0.441
CRP, mean (SD)	8.2 (12.9)	7.6	9.6	0.211
Number of painful entheses, mean (SD)	4.1 (5.0)	3.9	4.6	0.205
BASFI, mean (SD)	29.7 (21.3)	29.4	30.5	0.686
Radiographic status				
Local reading positive mNY criteria, n (%)	58 (19.7%)	25 (12.1%)	33 (37.5%)	<0.001
Central reading positive mNY criteria, n (%)	44 (14.9%)	15 (7.2%)	29 (33.0%)	<0.001
mNY score, mean (SD)	1.0 (1.2)	0.6	2.1	<0.001
mSASSS score, mean (SD)	0.53 (1.1)	0.2	1.3	<0.001
MRI sacroiliitis, n (%)	112 (38.0%)	66 (31.9%)	46 (52.3%)	<0.001
MRI SIJ SPARCC score, mean (SD)	3.2 (6.0)	2.3	5.3	<0.001
MRI Spine SPARCC score, mean (SD)	2.6 (6.1)	1.4	5.4	<0.001
Treatments				
NSAIDs consumption, n (%)	277 (93.9%)	198 (95.7%)	79 (89.8%)	0.662
Responsiveness to NSAID	261 (88.5%)	185 (89.4%)	76 (86.4%)	0.748

BMI: Body Mass Index; BASDAI: Bath Ankylosing Spondylitis Disease Activity Index; ASDAS: Ankylosing Spondylitis Disease Activity Score; CRP: C-reactive protein; BASFI: Bath Ankylosing Spondylitis Functional Index; mSASSS: modified Stoke Ankylosing Spondylitis Spinal Score; mNY: modified New-York; MRI: magnetic resonance imaging; NSAID: non-steroidal anti-inflammatory drug; SPARCC: SPondyloArthritis Research Consortium of Canada.

Results

Description and bivariate analyses

Baseline characteristics of patients from the DESIR cohort have previously been described (29), and the data of the 295 patients with no missing data regarding the outcome and the variables of interest (see below) are summarised in Table I.

Among the 295 patients included, 88 (29.8%) were considered as progressors according to our definition. Among the progressors: 46 (52.3%) were progressors at the spine level, 33 (37.5%) at the SIJ level, and 22 (25.0%) at both the spine and SIJ. All further analyses are based on these 295 patients.

The bivariate analyses evaluating the level of association between radiographic progression and each potential baseline variable of interest retained the following variables: age, gender, smoking status, profession (white collar *vs.* blue collar job), history of enthesitis, history of dactylitis, CRP, number of painful entheses, mNY criteria, mNY score, mSASSS score, MRI sacroiliitis, MRI SIJ SPondyloArthritis Research Consortium of Canada (SPARCC) score and MRI Spine SPARCC score.

Traditional models

- Logistic regression model

We performed a regression logistic

Table II. Multiple regression logistic model (M1 model).

Patient characteristics	*p*-value	OR (CI 95%)
Age	0.229	1.02 (0.99 – 1.06)
Gender	0.680	1.15 (0.58 – 2.29)
Smoking status	0.033	**1.94 (1.05 – 3.56)**
Profession	0.682	1.24 (0.44 – 3.39)
Disease general characteristics		
History of enthesitis	0.226	0.68 (0.36 – 1.27)
History of dactylitis	0.775	0.87 (0.32 – 2.15)
Disease activity		
CRP	0.603	1.01 (0.98 – 1.03)
Number of painful entheses	0.020	**1.07 (1.01 – 1.14)**
Imaging		
mNY score	<0.001	**2.14 (1.44 – 3.25)**
mSASSS score	0.006	**1.48 (1.14 – 1.99)**
Local reading mNY criteria	0.883	0.93 (0.35 – 2.36)
Central reading mNY criteria	0.227	0.42 (0.10 – 1.70)
MRI sacroiliitis	0.569	1.24 (0.58 – 2.60)
MRI SIJ SPARCC score	0.866	1.00 (0.95 – 1.06)
MRI Spine SPARCC score	0.116	0.96 (0.91 – 1.02)

OR: Odds ratio; CI95%: 95% confidence intervals; CRP: C-reactive protein; mNY: modified New-York; MRI: magnetic resonance imaging; NY: New-York, mSASSS: modified Stoke Ankylosing Spondylitis Spinal Score; SPARCC: SpondyloArthritis Research Consortium of Canada. n=295.

Table III. Comparison of 10-fold cross-validated AUC between traditional and machine learning models.

Models	Cross-validated AUC
Traditional models	
M2 (step AIC method)	0.79
M3 (LASSO method)	0.78
Machine learning approach	
SL Discrete Bayesian Additive Regression Trees Samplers (DBARTS)	0.76
SL Generalized Additive Models (GAM)	0.77
SuperLearner (SL)	0.74
SL LASSO	0.73
SL Random Forest	0.71
SL stepAIC	0.69
SL Multivariate adaptive polynomial spline regression (polymars)	0.68
SL Recursive Partitioning And Regression Trees (RPART)	0.62

AUC: area under the curve; AIC: Akaike information criterion; LASSO: least absolute shrinkage and selection Operator; SL: SuperLearner. n=295.

model, to explain the outcome of radiographic progression, with the 14 selected variables. Odds ratio (OR) and *p*-values are given Table II. In this model (M1), 4 variables were significantly associated with radiographic progression: smoking status (OR (IC95%)=1.94 (1.05–3.56), *p*=0.03), number of painful entheses on clinical examination (OR (IC95%)=1.07 (1.01–1.14), *p*=0.02, mNY score (OR per point (IC95%)=2.14 (1.44–3.25), *p*<0.001), mSASSS score (OR per point (IC95%)=1.48 (1.14–1.99), *p*=0.006). The 10-fold cross-validated AUC was 0.75. Brier score was 0.172 and Hosmer-Lemeshow goodness of fit was *p*=0.65.

- StepAIC method

First, we used the stepAIC method, which selected 4 variables in the final model (M2): smoking status, number of painful entheses on clinical examination, mNY score, mSASSS score.

For this model the 10-fold cross-validated AUC was 0.79, Brier score was 0.160 and Hosmer-Lemeshow goodness of fit was *p*=0.13.

- LASSO method

After applying LASSO, 10 variables were removed from M1. The 5 remaining variables were the same as in M2 model (smoking status, number of painful entheses on clinical examination, mNY score and mSASSS score) alongside with age. The final model (M3)'s 10-fold cross-validated AUC was 0.78. Brier score was 0.159 and Hosmer-Lemeshow goodness of fit was *p*=0.35.

- Sensitivity analysis: manual backward method after multiple imputation of missing data

After applying manual backward selection of variables on the whole population (after multiple imputation of missing data), we got a model (M4) that had only 3 explanatory variables: mNY score, mSASSS score, MRI SIJ SPARCC score. For this model, 10-fold cross-validated AUC was 0.75, Brier score was 0.168 and Hosmer-Lemeshow goodness of fit was *p*=0.63. Two variables were consistently found in all traditional models: baseline mNY score and mSASSS score. The most accurate traditional model was the M2 model (obtained with stepAIC method). Indeed, it had the highest 10-fold cross-validated AUC and the lowest AIC. Detailed traditional models are presented in Supplementary Table S2.

Machine Learning approach

The GAM, the DBARTS and the Super Learner models were the most accurate models with 10-fold cv-AUC of: 0.77, 0.76 and 0.74, respectively (Table III).

Comparison between traditional models and ML approach

The accuracy of the traditional models was compared to ML approach, based on 10-fold cross-validated AUC (Table III and Fig. 1). The two best traditional performing models were the M2 model (stepAIC method) with a cross-validated AUC of 0.79 and the M3 model (LASSO method) with a cross-validated AUC of 0.78. The two best models from ML approach gave comparable results: the GAM with a cross-validated AUC of 0.77 and the DBARTS with a cross-validated AUC of 0.76.

Discussion

Our analyses showed similar accuracy for traditional models (best cv-AUC = 0.79) compared to ML models (best cv-AUC = 0.77) in the DESIR cohort. Among the ML methods with the best

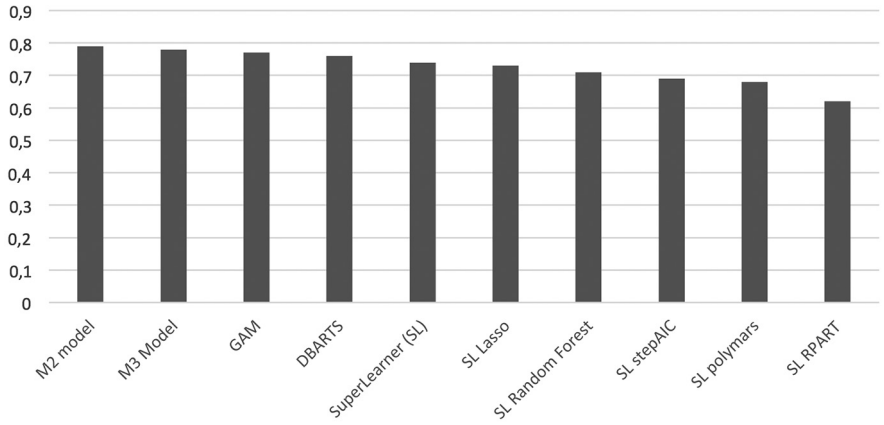

Fig. 1. Comparison of 10-fold cross-validated AUC between traditional and machine learning models.

accuracy were DBARTS and GAM. It is worth noting that prediction was very accurate with our traditional models and ML techniques, as cv-AUC of our four best models were >0.75.

Our study has some strengths. First, our analyses are based on the DESIR cohort which is the largest early axSpA cohort with extensive and longitudinal clinical and imaging data collection over 5 years (15). Besides, to the best of our knowledge, this is the first analysis aiming at comparing machine learning modelling to traditional models to evaluate radiographic progression in early axSpA.

Our study has some limits too. First, we did not confirm our hypothesis, *i.e.* that machine learning models would be more accurate than traditional models as suggested in some previous studies (10-12, 44). For instance, Jochems et al. were able to get a better AUC with machine learning random forest approach (AUC = 0.66 (95%CI: 0.54–0.77) compared to a traditional model (AUC = 0.55 (95%CI = 0.46–0.63)) for death prediction following radiotherapy in non-small cell lung cancer (11). In a systematic review on neurosurgical outcome prediction, Senders et al were able to show an improvement in AUC of 0.06 with ML compared to traditional logistic regression (12).

Another limit is that we decided to focus only on patients who would qualify as 'completers' for this analysis and who did not have any missing data on the variables of interest, resulting in 295 out of 708 patients. Nevertheless, when we performed our sensitivity analysis based on backward method after multiple imputations, results were comparable to those obtained with the 2 traditional models (stepAIC and LASSO), with the same two consistent predictive variables (mNY and mSASS at baseline).

Recently, ML has gained popularity but its definition is still a matter of debate and varies according to studies (45-46). It has been claimed that owing to its flexibility, ML would perform better than traditional statistical models (45-46). However, in 2019, a systematic review by Christodoulou *et al*. showed no advantage of machine learning over logistic regression for clinical prediction models, based on AUC, when comparisons had low risk of bias (45). The main advantage of ML might be that it can handle a huge amount of predictors but one of its pitfalls is that it probably needs more data than traditional models (45, 47-48). Indeed, it was shown in 2014 that some ML methods needed at least 10 times more events per variable than logistic regression (48). Furthermore another limit, regarding "SuperLearner" package on R, is that it is a "black box" algorithm making it very difficult for the user to fully understand the contribution of each covariate (49). Our traditional and ML gave models gave similar results and seem to accurately predict radiographic progression in these early axSpA patients, mainly by the presence of radiographic changes at baseline. Other studies involving other kind of data or other artificial intelligence methods (*i.e.* deep learning) might be necessary to obtain more accurate estimates of radiographic progression in early axSpA, particularly in the subgroup of patient without any radiographic involvement at baseline.

Acknowledgments

The DESIR study is conducted with Assistance Publique Hôpitaux de Paris as the sponsor. The DESIR study is also under the umbrella of the French Society of Rheumatology, which financially supports the cohort. An unrestricted grant from Pfizer has been allocated for the first 10 years. The DESIR cohort is conducted under the control of Assistance Publique Hôpitaux de Paris via the Clinical Research Unit Paris Centre and under the umbrella of the French Society of Rheumatology and Institut national de la santé et de la recherche médicale (Inserm). Database management is performed within the Department of Epidemiology and Biostatistics (Dr Pascale Fabbro-Peray, D.I.M., and Nîmes, France). We also wish to thank the different regional participating centres: Pr Maxime Dougados, Dr Anna Moltó (Paris-Cochin), Pr Philippe Dieudé (Paris-Bichat), Pr Laure Gossec (Paris-La Pitie-Salpétriere), Pr Francis Berenbaum (Paris-Saint- Antoine), Pr Pascal Claudepierre (Creteil), Pr Maxime Breban (Boulogne- Billancourt), Dr Bernadette Saint-Marcoux (Aulnay-sous-Bois), Pr Philippe Goupille (Tours), Pr Jean Francis Maillefert (Dijon), Dr Emmanuelle Dernis (Le Mans), Pr Daniel Wendling (Besancon), Pr Bernard Combe (Montpellier), Pr Liana Euller-Ziegler (Nice), Pr Pascal Richette (Paris Lariboisière), Pr Pierre Lafforgue (Marseille), Pr Patrice Fardellone, Dr Patrick Boumier (Amiens), Pr Martin Soubrier (Clermont-Ferrand), Dr Nadia Mehsen (Bordeaux), Pr Damien Loeuille (Nancy), Pr Rene-Marc Flipo (Lille), Pr Alain Saraux (Brest), Dr Stephan Pavy (Le Kremlin-Bicetre), Pr Adeline Ruyssen-Witrand (Toulouse), Pr Olivier Vittecoq (Rouen). We wish to thank the research nurses, the staff members of the Clinical Research Unit of Paris Centre, the staff members of the Biological Resource Centre of Bichat Hospital, the staff members of the Department of Statistics of Nımes and all the investi-

gators, and in particular Jerome Allain, Thierry Lequerre, Beatrice Banneville, Julien Champey, Christine Piroth, Anne Tournadre, Sophie Trijau, Salah Ferkal, Clement Prati, Marie-Agnes Timsit, Eric Toussirot for active patient recruitment and monitoring.

References

1. DOUGADOS M, BAETEN D: Spondyloarthritis. *Lancet* 2011; 377: 2127-37. https://doi.org/10.1016/s0140-6736(11)60071-8
2. CARDELLI C, MONTI S, TERENZI R, CARLI L: One year in review 2021: axial spondyloarthritis. *Clin Exp Rheumatol* 2021; 39: 1272-81. https://doi.org/10.55563/clinexprheumatol/jlyd1l
3. RAMIRO S, STOLWIJK C, VAN TUBERGEN A *et al*.: Evolution of radiographic damage in ankylosing spondylitis: a 12 year prospective follow-up of the OASIS study. *Ann Rheum Dis* 2015; 74: 52-9. https://doi.org/10.1136/annrheumdis-2013-204055
4. PODDUBNYY D, HAIBEL H, LISTING J *et al*.: Baseline radiographic damage, elevated acute-phase reactant levels, and cigarette smoking status predict spinal radiographic progression in early axial spondylarthritis. *Arthritis Rheum* 2012; 64: 1388-98. https://doi.org/10.1002/art.33465
5. DOUGADOS M, SEPRIANO A, MOLTO A *et al*.: Sacroiliac radiographic progression in recent onset axial spondyloarthritis: the 5-year data of the DESIR cohort. *Ann Rheum Dis* 2017; 76: 1823-8. https://doi.org/10.1136/annrheumdis-2017-211596
6. DOUGADOS M, DEMATTEI C, VAN DEN BERG R *et al*.: Rate and predisposing factors for sacroiliac joint radiographic progression after a two-year follow-up period in recent-onset spondyloarthritis: radiographic progression in the sacroiliac joints in axial SpA. *Arthritis Rheumatol* 2016; 68: 1904-13. https://doi.org/10.1002/art.39666
7. BLACHIER M, CANOUI-POITRINE F, DOUGADOS M *et al*.: Factors associated with radiographic lesions in early axial spondyloarthritis. Results from the DESIR cohort. *Rheumatology* 2013; 52: 1686-93. https://doi.org/10.1093/rheumatology/ket207
8. PEDUZZI P, CONCATO J, KEMPER E, HOLFORD TR, FEINSTEIN AR: A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; 49: 1373-9. https://doi.org/10.1016/s0895-4356(96)00236-3
9. RANGANATHAN P, PRAMESH CS, AGGARWAL R: Common pitfalls in statistical analysis: logistic regression. *Perspect Clin Res* 2017; 8: 148-51. https://doi.org/10.4103/picr.picr_87_17
10. CHURPEK MM, YUEN TC, WINSLOW C, MELTZER DO, KATTAN MW, EDELSON DP: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards: *Crit Care Med* 2016; 44: 368-74. https://doi.org/10.1097/ccm.0000000000001571
11. JOCHEMS A, EL-NAQA I, KESSLER M *et al*.: A prediction model for early death in non-small cell lung cancer patients following curative-intent chemoradiotherapy. *Acta Oncol* 2018; 57: 226-30. https://doi.org/10.1080/0284186x.2017.1385842
12. SENDERS JT, STAPLES PC, KARHADE AV *et al*.: Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg* 2018; 109: 476-86.e1. https://doi.org/10.1016/j.wneu.2017.09.149
13. LEZCANO-VALVERDE JM, SALAZAR F, LEÓN L *et al*.: Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach. *Sci Rep* 2017; 7. https://doi.org/10.1038/s41598-017-10558-w
14. DEODHAR A, ROZYCKI M, GARGES C *et al*.: Use of machine learning techniques in the development and refinement of a predictive model for early diagnosis of ankylosing spondylitis. *Clin Rheumatol* 2020; 39(4): 975-82. https://doi.org/10.1007/s10067-019-04553-x
15. DOUGADOS M, ETCHETO A, MOLTO A *et al*.: Clinical presentation of patients suffering from recent onset chronic inflammatory back pain suggestive of spondyloarthritis: The DESIR cohort. *Joint Bone Spine* 2015; 82: 345-51. https://doi.org/10.1016/j.jbspin.2015.02.006
16. CREEMERS MC, FRANSSEN MJ, VAN'T HOF MA, GRIBNAU FW, VAN DE PUTTE LB, VAN RIEL PL: Assessment of outcome in ankylosing spondylitis: an extended radiographic scoring system. *Ann Rheum Dis* 2005; 64: 127-9. https://doi.org/10.1136/ard.2004.020503
17. RAMIRO S, CLAUDEPIERRE P, SEPRIANO A *et al*.: Which scoring method depicts spinal radiographic damage in early axial spondyloarthritis best? Five-year results from the DESIR cohort. *Rheumatology* (Oxford) 2018; 57: 1991-2000. https://doi.org/10.1093/rheumatology/key185
18. VAN DER LINDEN S, VALKENBURG HA, CATS A: Evaluation of diagnostic criteria for ankylosing spondylitis. *Arthritis Rheum* 1984; 27: 361-8. https://doi.org/10.1002/art.1780270401
19. BARALIAKOS X, HAIBEL H, LISTING J, SIEPER J, BRAUN J: Continuous long-term anti-TNF therapy does not lead to an increase in the rate of new bone formation over 8 years in patients with ankylosing spondylitis. *Ann Rheum Dis* 2014; 73: 710-5. https://doi.org/10.1136/annrheumdis-2012-202698
20. FAY MP, PROSCHAN MA: Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Stat Surv* 2010; 4: 1-39. https://doi.org/10.1214/09-ss051
21. DU PREL J-B, RÖHRIG B, HOMMEL G, BLETTNER M: Choosing statistical tests: part 12 of a series on evaluation of scientific publications. *Dtsch Aerztebl Int* 2010; 107: 343-8. https://doi.org/10.3238/arztebl.2010.0343
22. KIM H-Y: Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restor Dent Endod* 2017; 42: 152. https://doi.org/10.5395/rde.2017.42.2.152
23. HARRELL FE JR: Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York, NY: Springer New York; 2001.
24. ZHANG Z: Variable selection with stepwise and best subset approaches. *Ann Transl Med* 2016; 4: 136. https://doi.org/10.21037/atm.2016.03.35
25. COLLIGNON O, HAN J, AN H, OH S, LEE Y: Comparison of the modified unbounded penalty and the LASSO to select predictive genes of response to chemotherapy in breast cancer. *PLoS One* 2018; 13(10): e0204897. https://doi.org/10.1371/journal.pone.0204897
26. JUNG Y, HU J: A K-fold averaging cross-validation procedure. *J Nonparametric Stat* 2015; 27: 167-79. https://doi.org/10.1080/10485252.2015.1010532
27. ROBIN X, TURCK N, HAINARD A *et al*.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; 12. https://doi.org/10.1186/1471-2105-12-77
28. VAN DER LAAN MJ, POLLEY EC, HUBBARD AE: Super learner. *Stat Appl Genet Mol Biol* 2007; 6: Article 25. https://doi.org/10.2202/1544-6115.1309
29. MARK J. VAN DER LAAN ECP: Super Learner In Prediction. Berkeley Div Biostat Work Pap Ser.
30. CHEN X, WANG M, ZHANG H: The use of classification trees for bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov* 2011; 1(1): 55-63. https://doi.org/10.1002/widm.14
31. COOPER JN, MINNECI PC, DEANS KJ: Postoperative neonatal mortality prediction using superlearning. *J Surg Res* 2018; 221: 311-9. https://doi.org/10.1016/j.jss.2017.09.002
32. GIACOMELLI I, JHA S, KLEIMAN R, PAGE D, YOON K: Privacy-Preserving Collaborative Prediction using Random Forests. *AMIA Jt Summits Transl Sci Proc* 2019; 2019: 248-57. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc6568057/
33. ZHAO Y, ZHENG W, ZHUO DY *et al*.: Bayesian additive decision trees of biomarker by treatment interactions for predictive biomarker detection and subgroup identification. *J Biopharm Stat* 2018; 28(3): 534-49. https://doi.org/10.1080/10543406.2017.1372770
34. HASTIE T, TIBSHIRANI R: Generalized additive models for medical research. *Stat Methods Med Res* 1995; 4: 187-96. https://doi.org/10.1177/096228029500400302
35. RABIDEAU DJ, PEI PP, WALENSKY RP, ZHENG A, PARKER RA: Implementing generalized additive models to estimate the expected value of sample information in a microsimulation model: results of three case studies. *Med Decis Making* 2018; 38: 189-99. https://doi.org/10.1177/0272989x17732973
36. FRIEDMAN JH: Multivariate adaptive regression splines. *Ann Statist* 1991; 19: 1-67. https://doi.org/10.1214/aos/1176347963
37. FRIEDMAN JH, ROOSEN CB: An introduction to multivariate adaptive regression splines. *Stat Methods Med Res* 1995; 4: 197-217. https://doi.org/10.1177/096228029500400303
38. BAKER SG: The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J Natl Cancer Inst* 2003; 95(7): 511-5. https://doi.org/10.1093/jnci/95.7.511
39. WU Y-C, LEE W-C: Alternative performance measures for prediction models. *PLoS One* 2014; 9(3): e91249. https://doi.org/10.1371/journal.pone.0091249
40. PAUL P, PENNELL ML, LEMESHOW S: Standardizing the power of the Hosmer-Lemeshow

goodness of fit test in large data sets. *Stat Med* 2013; 32(1): 67-80. https://doi.org/10.1002/sim.5525

41. KRAMER AA, ZIMMERMAN JE: Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med* 2007; 35(9): 2052-6. https://doi.org/10.1097/01.ccm.0000275267.64078.b0
42. LEMESHOW S, HOSMER DW: A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol* 1982; 115(1): 92-106. https://doi.org/10.1093/oxfordjournals.aje.a113284
43. AZUR MJ, STUART EA, FRANGAKIS C, LEAF PJ: Multiple imputation by chained equations: what is it and how does it work?: Multiple imputation by chained equations. *Int J Methods Psychiatr Res* 2011; 20(1) https://doi.org/10.1002/mpr.329: 40-9.
44. HOLLON TC, PARIKH A, PANDIAN B *et al*.: A machine learning approach to predict early outcomes after pituitary adenoma surgery. *Neurosurg Focus* 2018; 45(5): E8. https://doi.org/10.3171/2018.8.focus18268
45. CHRISTODOULOU E, MA J, COLLINS GS, STEYERBERG EW, VERBAKEL JY, VAN CALSTER B: A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; 110: 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004
46. LUO W, PHUNG D, TRAN T *et al*.: Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016; 18(12): e323. https://doi.org/10.2196/jmir.5870
47. DEO RC, NALLAMOTHU BK: Learning about machine learning: the promise and pitfalls of big data and the electronic health record. *Circ Cardiovasc Qual Outcomes* 2016; 9(6): 618-20.https://doi.org/10.1161/circoutcomes. 116.003308
48. VAN DER PLOEG T, AUSTIN PC, STEYERBERG EW: Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014; 14. https://doi.org/10.1186/1471-2288-14-137
49. NAIMI AI, BALZER LB: Stacked generalization: an introduction to super learning. *Eur J Epidemiol* 2018; 33(5): 459-64. https://doi.org/10.1007/s10654-018-0390-z