International Journal of Medical Informatics 110 (2018) 31–41

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Evaluation of three machine learning models for self-referral decision support on low back pain in primary care

Wendy Oude Nijeweme-d'Hollosy[a,*], Lex van Velsen[a,b], Mannes Poel[c], Catharina G.M. Groothuis-Oudshoorn[d], Remko Soer[e,f], Hermie Hermens[a,b]

[a] University of Twente, CTIT, MIRA, EWI/BSS Telemedicine, Enschede, The Netherlands
[b] Roessingh Research and Development, Telemedicine cluster, Enschede, The Netherlands
[c] University of Twente, EWI/Human Media Interaction, The Netherlands
[d] MIRA, Health Technology and Services Research University of Twente, Enschede, The Netherlands
[e] University of Groningen, University Medical Centre Groningen, Groningen Spine Centre, Groningen, The Netherlands
[f] Saxion University of Applied Science, Enschede, The Netherlands

ARTICLE INFO

Keywords:
Clinical decision support system
Low back pain
Machine learning
Primary care
Self-referral

ABSTRACT

Background: Most people experience low back pain (LBP) at least once in their life and for some patients this evolves into a chronic condition. One way to prevent acute LBP from transiting into chronic LBP, is to ensure that patients receive the right interventions at the right moment. We started research in the design of a clinical decision support system (CDSS) to support patients with LBP in their self-referral to primary care. For this, we explored the possibilities of using supervised machine learning. We compared the performances of the three classification models − i.e. 1. decision tree, 2. random forest, and 3. boosted tree − to get insight in which model performs best and whether it is already acceptable to use this model in real practice.
Methods: The three models were generated by means of supervised machine learning with 70% of a training dataset (1288 cases with 65% GP, 33% physio, 2% self-care cases). The cases in the training dataset were fictive cases on low back pain collected during a vignette study with primary healthcare professionals. We also wanted to know the performance of the models on real-life low back pain cases that were not used to train the models. Therefore we also collected real-life cases on low back pain as test dataset. These cases were collected with the help of patients and healthcare professionals in primary care. For each model, the performance was measured during model validation − with 30% of the training dataset −as well as during model testing − with the test dataset containing real-life cases. The total observed accuracy as well as the kappa, and the sensitivity, specificity, and precision were used as performance measures to compare the models.
Results: For the training dataset, the total observed accuracies of the decision tree, the random forest and boosted tree model were 70%, 69%, and 72% respectively. For the test dataset, the total observed accuracies were 71%, 53%, and 71% respectively. The boosted tree appeared to be the best for predicting a referral advice with a fair accuracy (Kappa between 0.2 and 0.4). Next to this, the measured evaluation measures show that all models provided a referral advice better than just a random guess. This means that all models learned some implicit knowledge of the provided referral advices in the training dataset.
Conclusions: The study showed promising results on the possibility of using machine learning in the design of our CDSS. The boosted tree model performed best on the classification of low back pain cases, but still has to be improved. Therefore, new cases have to be collected, especially cases that are classified as self-care cases. This to be sure that also the self-care advice can be predicted well by the model.

1. Background

Most people experience low back pain (LBP) at least once in their life. As such, it is one of the most common health problems in the world [1–3]. A formal definition of LBP is "*pain, muscle tension, or stiffness localized below the costal margin and above the inferior gluteal folds, with or*

Abbreviations: CDSS, computerized decision support system; C.I., confidence interval; COPD, chronic obstructive pulmonary disease; GP, general practitioner; LBP, low back pain; RF, random forest

[*] Corresponding author at: University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands.
E-mail address: w.dhollosy@utwente.nl (W. Oude Nijeweme-d'Hollosy).

https://doi.org/10.1016/j.ijmedinf.2017.11.010
Received 4 August 2017; Received in revised form 9 November 2017; Accepted 19 November 2017

without leg pain [4]". This means that LBP is in fact a symptom referring to the location of the problem, rather than a specific pathology that causes the problem [5].

Some patients with LBP develop a chronic condition. The risk of chronic LBP continues to increase with age [3,6]. Because LBP causes considerable disability and financial burden globally [7], it is of importance to prevent the development of chronic LBP wherever possible. One way to prevent acute LBP from transiting into chronic LBP, is to ensure that patients receive the right interventions at the right moment [8]. However, this group of patients is heterogeneous, and individual patients respond differently to interventions. Therefore, relevant studies have been conducted in an attempt to classify patients with LBP to the most optimal interventions [9–13].

Normally, a patient with a new episode of LBP starts in primary care [11] by visiting a general practitioner (GP) or physiotherapist. In an increasing number of countries, patients with musculoskeletal disorders can make use of patient self-referral to a physiotherapist [14,15]. Characteristics of patients that utilize self-referral are higher education level, a shorter duration of symptoms and recurrent symptoms [16,17]. However, for a group of patients it is still unclear what to do first: consult a GP or consult a physiotherapist. There is also a third option, namely performing self-care at first [18]. During self-care, the patient is not treated by a professional and continues ordinary activities within the limits permitted by the pain. This usually leads to faster recovery than either bed rest or back-mobilizing exercises [14]. When a patient visits a GP or physiotherapist, (s)he can refer the patient further to other options when needed. In the Netherlands, for example, the GP can refer the patient to the emergency room, but also to other secondary and tertiary care specialists as neurology, orthopedics, spine centers, pain centers, or psychologically augmented physiotherapy in the case of psychological and social factors causing the LBP[12]. In this paper, we focused on self-referral to GP, physiotherapist, or self-care as these are the first steps in a new episode of LBP in the Dutch care system, and further referral to other options sought by patients experiencing LBP can only be taken if one or more of these three steps have been performed.

In 2015, we started research to design a clinical decision support system (CDSS) to support patients with LBP in their self-referral process [19]. This is a classification process that leads to one of the three following referral advices: 1. consult a GP, 2. consult a physiotherapist, or 3. perform self-care. As self-referral can be seen as a classification process, we opted for supervised machine learning to design a classification model representing this process. Machine learning offers algorithms that can be used to learn computers based on data [20]. In supervised machine learning, a classification model learns from labelled examples.

Machine learning is increasingly used in healthcare informatics [21], also in the case of patient referral. Recent examples are systems in emergency departments to identify patients with suspected infection [22] and to identify low-complexity patients that can be included in a separate fast track patient stream to save waiting time and capacity [23]. In case of musculoskeletal problems, the Work Assessment Triage Tool (WATT) is an example of a machine learned CDSS that refers workers with musculoskeletal injuries to optimal rehabilitation interventions [24]. For LBP in particular, there is the Nijmegen Decision Tool for referral of chronic LBP to be used by secondary or tertiary spine care specialists [5]. However, the design of this tool was not based on a machine learning approach and is not intended for patient self-referral in primary care.

In this paper, we explore the possibilities of using supervised machine learning in the design of our CDSS to support patients with LBP in their self-referral to primary care, as machine learning can often be successfully applied for classification problems [25]. Our exploration is the follow-up of two steps we already have undertaken so far: 1. an inventory of important features to classify LBP [19], and 2. a vignette study in which fictive cases of LBP were judged on referral advice by healthcare professionals [26]. The vignette study has resulted in a dataset containing labelled examples that can be used for supervised machine learning. In this paper, this dataset is used as *training dataset* to train three machine learning models, i.e. 1. decision tree, 2. random forest, and 3. boosted tree. Next to this, we also describe the process used to construct a *test dataset* with real-life cases of LBP. With this test dataset, we compare the performances of the three classification models on real-life cases. In this way, we get insight in which model performs best and whether it is already acceptable to use this model in real practice.

2. Methods

2.1. Machine learning

At first, the intension was to build a decision tree only, as decision trees are self-explanatory and easy to follow [28]. However, a decision tree is a single classifier and ensembles of classifiers often perform better than a single classifier [29]. Therefore, we also focused on tree ensembles. The following three classification models were generated 1. decision tree, 2. random forest, and 3. boosted tree. The first model is a single tree [30], the second and third models are ensembles of trees. In a random forest, different decision trees are generated on subsets that are sampled with replacement from the original training dataset. Classification of a new case takes place by majority vote of the trees in the random forest [31]. The difference with boosted tree is that for boosted tree the distribution of the training set for generating the next tree is adaptively changed, based on the performance of previous classifiers [32]. R [33] in RStudio [34] was used to train these classification models with our training dataset.

2.2. Datasets

2.2.1. Training dataset

The training dataset consisted of 1288 fictive cases of LBP. These cases were judged by 63 physiotherapists and GPs on referral advice during a vignette study [26]. Table 1 provides a detailed overview of the variables − 15 input variables, 1 response variable − that describe the cases in this training dataset. During the vignette study, cases of LBP were presented that were generated by using a fixed text in which the values of the 15 input variables varied randomly. No combination of variable values was used twice, therefore the training dataset exists of unique cases. The referral advices among the cases in the training dataset were classified as follows: 843 (65%) GP advice, 425 (33%) physiotherapy advice, and 20 (2%) self-care advice.

2.2.2. Test dataset

From September 2016 to April 2017, we collected a set of real-life cases of LBP to construct a test dataset. The intention was to collect as much as possible patient cases during the time the study was conducted. This was done in collaboration with 5 centres for physical therapy and 6 GP centres. We presented our study to the medical ethical committee. We received a statement that ethics approval was not required for our study, as the normal healthcare path was not influenced and the patients remained anonymous to the researchers.

The study design that was used to collect real-life cases of LBP is shown in Fig. 1. This process started when a patient with a new episode of LBP called a centre to make an appointment (1). Subsequently, the patient was asked to participate the study (2). If agreed, the patient was informed about the study and received a web-address to an online questionnaire with questions related to the input variables of Table 1. After the patient had given informed consent, the patient answered the questions (3). Next, the patient visited the healthcare professional of his/her preference. After the consult, the healthcare professional filled in a form indicating what the advice to the patient should have been: visit a GP, a physiotherapist, or perform self-care (4). The answers of

Table 1
Overview of the input variables (features) and response variable (output) that describe the cases in the dataset that was used to train the three classification models.

Name	Description	Type	Values
Age	The age of the patient	Input variable − Factor w/2 levels	“ < 50”, “ > = 50”
Wellbeing	The state of being healthy as perceived by the patient by using the questions of the WHO-5 Well-Being Index [27]	Input variable − Factor w/3 levels	“bad”, “good”, ”medium”
Course	The duration of the current low back pain episode	Input variable − Factor w/2 levels	“ < 2weeks”, “ > = 2weeks”
Analgesics	Does the patient use analgesics − e.g. paracetamol, ibuprofen or diclofenac − on a daily basis?	Input variable − Factor w/2 levels	“no”, “yes”
Trauma	Was the low back pain caused by a trauma?	Input variable − Factor w/2 levels	“no”, “yes”
Corticosteroid	Does the patient use corticosteroids − e.g. prednisone − on a daily basis?	Input variable − Factor w/2 levels	“no”, “yes”
Serious diseases	Does the patient has serious diseases, namely one of the following: osteoporosis, vertebral fracture, cancer, rheumatic disorder (e.g., Bechterew disease, osteoarthritis), narrowing of the spinal canal (Canal Stenosis), shifted vertebra (s) or damaged vertebrae demonstrated on X-rays?	Input variable − Factor w/2 levels	“no”, “yes”
Weigthloss > 5 kg	Has the patient lost more than 5 kg in the past month without a reason e.g. a diet?	Input variable − Factor w/2 levels	“no”, “yes”
Continouspain	Does the patient currently has constant pain, which does not decrease with rest or when changing posture?	Input variable − Factor w/2 levels	“no”, “yes”
Nocturnalpain	Does the patient also has low back pain at night that wakes the patient up?	Input variable − Factor w/2 levels	“no”, “yes”
Neurogenicsignals	Does the patient has more pain if the patient has to cough or sneeze, or when the patient is lifting something?	Input variable − Factor w/2 levels	“no”, “yes”
Radiation	Does the patient suffer from tingling or pangs in one or both legs?	Input variable − Factor w/2 levels	“no”, “yes”
Lossmusclestrength	Does the patient has reduced strength in one or both legs?	Input variable − Factor w/2 levels	“no”, “yes”
Failuresymptoms	Does the patient suffer from failure symptoms in one or both legs, which makes it impossible to move a leg, or legs, or leads to urinary loss?	Input variable − Factor w/2 levels	“no”, “yes”
Preference	Referral preference of the patient	Input variable − Factor w/3 levels	“GP”, “Physio”,”Selfcare”
Advice	Referral advice for this patient case	Response variable − Factor w/3 levels	“GP”, “Physio”,”Selfcare”

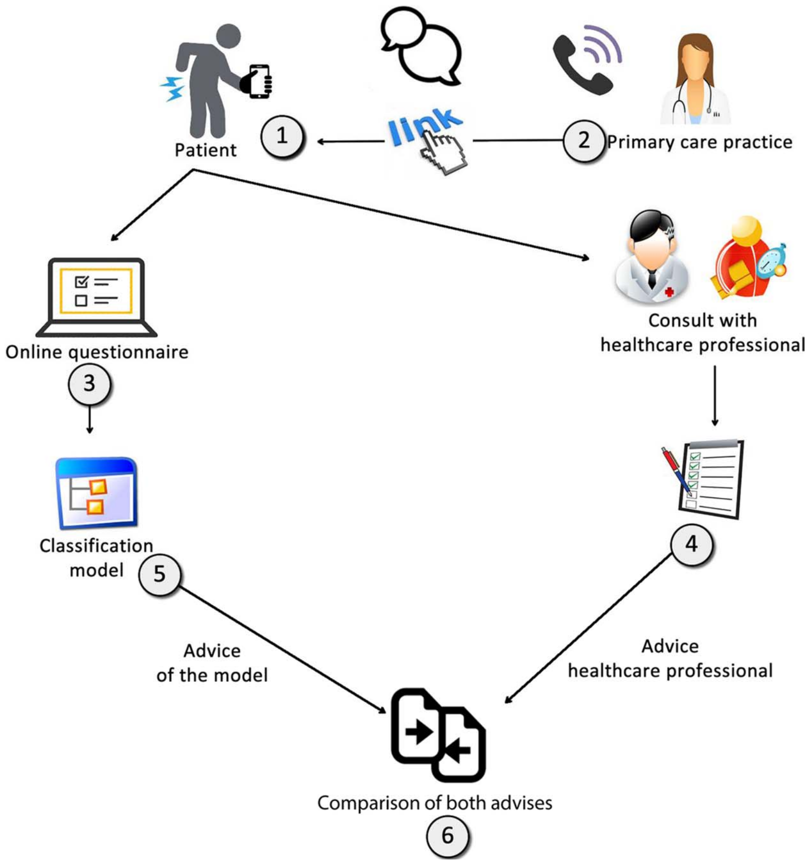
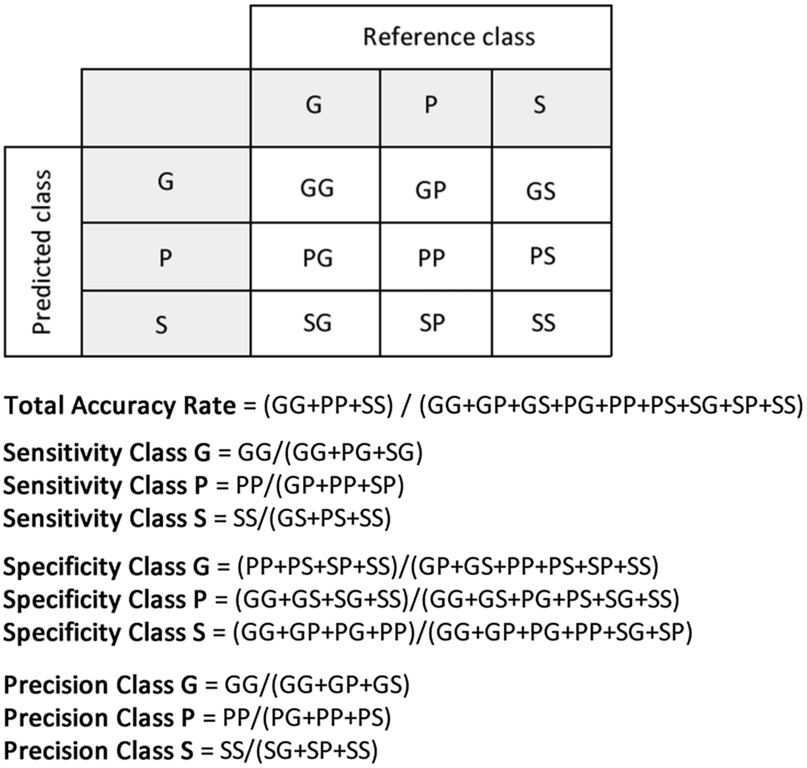


Fig. 1. Stduy design that was used to collect real-life cases on low back pain. These cases were used as test dataset in the evaluation of the three classification models.



		Reference class		
		G	P	S
Predicted class	G	GG	GP	GS
	P	PG	PP	PS
	S	SG	SP	SS

Total Accuracy Rate = (GG+PP+SS) / (GG+GP+GS+PG+PP+PS+SG+SP+SS)

Sensitivity Class G = GG/(GG+PG+SG)
Sensitivity Class P = PP/(GP+PP+SP)
Sensitivity Class S = SS/(GS+PS+SS)

Specificity Class G = (PP+PS+SP+SS)/(GP+GS+PP+PS+SP+SS)
Specificity Class P = (GG+GS+SG+SS)/(GG+GS+PG+PS+SG+SS)
Specificity Class S = (GG+GP+PG+PP)/(GG+GP+PG+PP+SG+SP)

Precision Class G = GG/(GG+GP+GS)
Precision Class P = PP/(PG+PP+PS)
Precision Class S = SS/(SG+SP+SS)

Fig. 2. Confusion matrix and evaluation measures that were used to explore the performances of the three classification models, where G represents the class GP, P the class physiotherapy, and S the class self-care.

the patients on the questions were entered into each classification model (5). Finally, per model the predicted advice was compared to the referral advice provided by the healthcare professional (6).

2.3. Model performance assessment

The models were explored by comparing their performances. A performance measure often used to evaluate a model is *accuracy*, also known as the *recognition rate*. However, using *accuracy* is only a good indicator in the evaluation of a model when the class distribution in the training dataset is well-balanced [20]. In our study, we had an unbalanced multiclass training dataset: 65% GP advice, 33% physiotherapy advice, and 2% self-care advice. Therefore, also other evaluation measures were taken into account (Fig. 2). Per model, we used a *confusion matrix* to calculate the *sensitivity* (true positive rate), the *specificity* (true negative rate), and the *precision* (positive predictive value) to gain more insight into the performances of the models. The kappa of the models were compared too. The kappa is a metrics for the strength of agreement of a model that compares the observed accuracy with the expected accuracy [35] with a Kappa of 0-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost perfect [35].

Each model was trained with 70% of the training dataset (model training), validated with 30% of the training dataset (model validation), and tested with the test dataset (model testing) (Fig. 3). The cases in the test dataset were not used to train the model to be able to measure the performances of the models more accurately [20]. For each model, all evaluation measures were calculated during model validation as well as during the model testing.

3. Results

3.1. Test dataset

In total, 45 patients completed the online questionnaire before visiting the healthcare professional. Next to this, 44 healthcare professionals provided a referral advice after seeing patients. However, not all 44 referral advices could be connected to a completed questionnaire, as some patients intended to participate into the study, but for some reason did not fill in the online questionnaire. Next to this, some patients filled in the questionnaire, but no referral advice was provided by the healthcare professional. In the end, 38 of the 45 completed questionnaires could be connected to a provided referral advice. This resulted into a set of 38 real-life cases of LBP.

The average age of the patients was 40.00 years (SD 14.53; range 17.00-79.00 years). Table 2 shows that 33 patients visited a physiotherapist and 5 visited a GP. The 38 cases were classified as follows: 4 (11%) GP advice, 30 (78%) physiotherapy advice, and 4 (11%) self-care advice. Thus the test dataset also became an unbalanced dataset. However, in contrast to the training dataset, in the test dataset physiotherapy advice was the overrepresented class. Table 2 shows that in the test dataset, just as in the training dataset, "self-care" was the underestimated class.

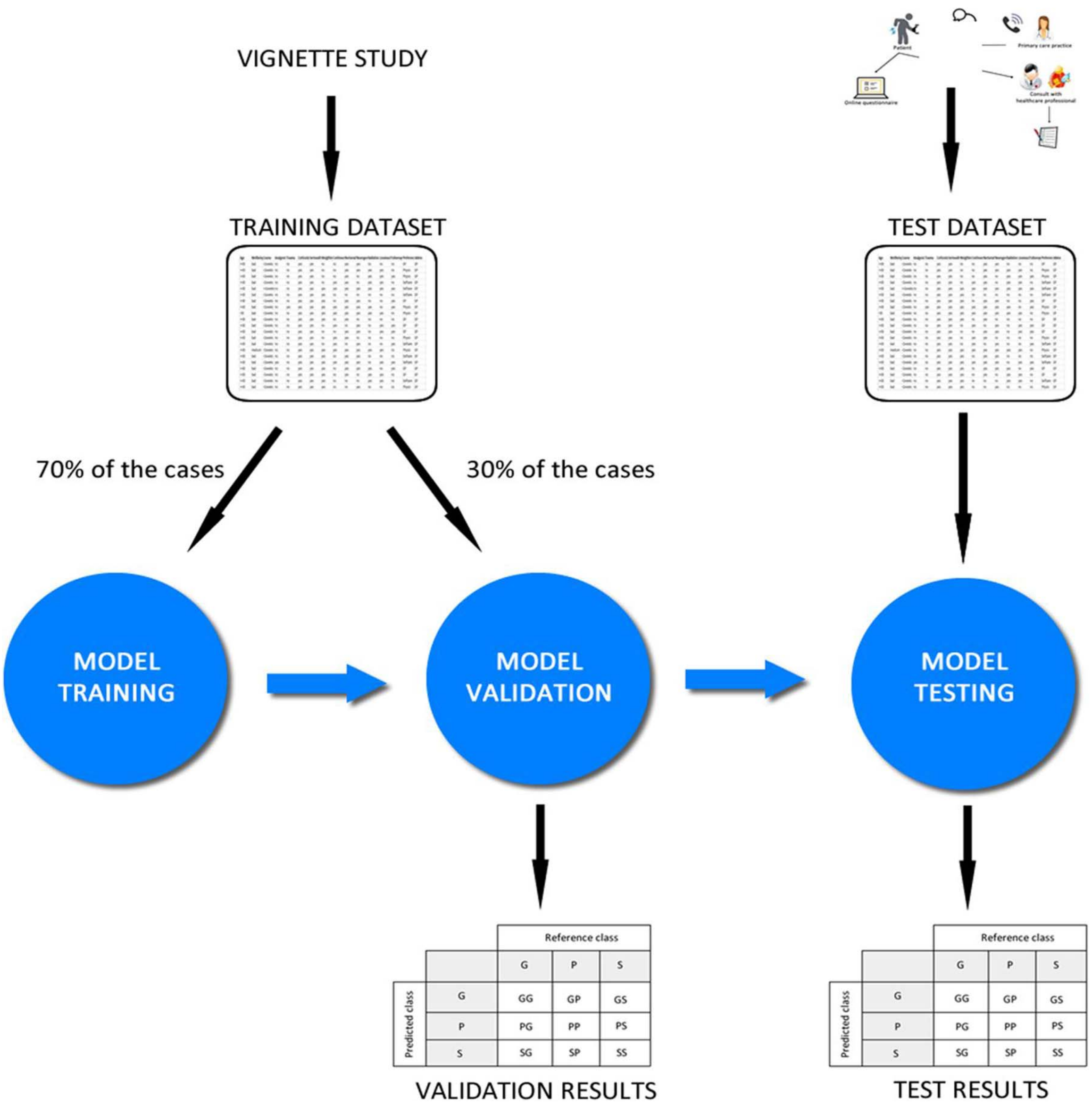
We asked the GPs in our study if they could explain why they did not see as many patients with LBP as the physiotherapists. It appeared that most patients with acute low back get advice from the doctors' assistant first on how to cope with the LBP and to wait a couple of days to see what happens in first instance. Then the GP did not see these patients. Next to this, the GPs also indicated that patients more often find the direct way to the physiotherapist for musculoskeletal problems.

By handling Table 2 as confusion matrix, we could determine the accuracy of the choice of a patient for a healthcare professional. We found a total accuracy rate of 0.868 (95% C.I. 0.719, 0.956). This means that in about 87% of the cases the patient consulted the same type of healthcare professional − GP or physiotherapist − as also was indicated in the referral advice.

3.2. Results of model training, model validation and model testing

3.2.1. Decision tree

The decision tree is shown in Fig. 4. This figure shows that from the original 15 features (Table 1) only 4 features were used in the decision



		Reference class		
		G	P	S
Predicted class	G	GG	GP	GS
	P	PG	PP	PS
	S	SG	SP	SS

		Reference class		
		G	P	S
Predicted class	G	GG	GP	GS
	P	PG	PP	PS
	S	SG	SP	SS

Fig. 3. Overview of the different phases in exploring the performance of each model as performed in this study.

Table 2
Overview of the numbers of healthcare professionals that were visited by the patients in the test dataset and the referral-advices as provided by these healthcare professionals. G represents the class GP, P the class physiotherapy, and S the class self-care.

		Provided advice healthcare professional			
Patient visited		**G**	**P**	**S**	**Total**
	G	3	1	0	**4**
	P	0	30	0	**30**
	S	2	2	0	**4**
	Total	**5**	**33**	**0**	**38**

nodes i.e. Weight loss, Wellbeing, Usage of corticosteroids, and Loss of muscle strength. Furthermore, this decision tree never provides a self-care advice, probably because only 2% of the cases in the training dataset was classified as self-care advice class and therefore never could reach the highest fraction of a class in a node of leaf.

Table 3 shows the *confusion matrix*, *accuracy*, *sensitivity, specificity*, and *precision* measures of the decision tree.

3.2.2. *Random forest*

A random forest cannot be presented like a decision tree, but Fig. 5 shows multiclass ROC curve of this random forest. For all three advice classes, the prediction performance of the random forest is better than just a random choice. Fig. 6 shows the determined variable importance in the random forest for each class. Weight Loss more than 5 kg is the most important feature in the process of classifying a referral advice.

Table 4 shows the *confusion matrix*, *accuracy*, *sensitivity, specificity*, and *precision* measures of the random forest.

3.2.3. *Boosted tree*

Fig. 7 shows the multiclass ROC curve of the boosted tree, which shows that for the boosted tree model the prediction performance is better than a random choice. Fig. 8 shows the determined total variable importance in the boosted tree. Again, Weight Loss more than 5 kg is the most important feature in the process of classifying a referral advice.

Table 5 shows the *confusion matrix*, *accuracy*, *sensitivity, specificity*, and *precision* measures of the boosted tree.

3.3. *Model comparison*

The measured performances show that all models provided a

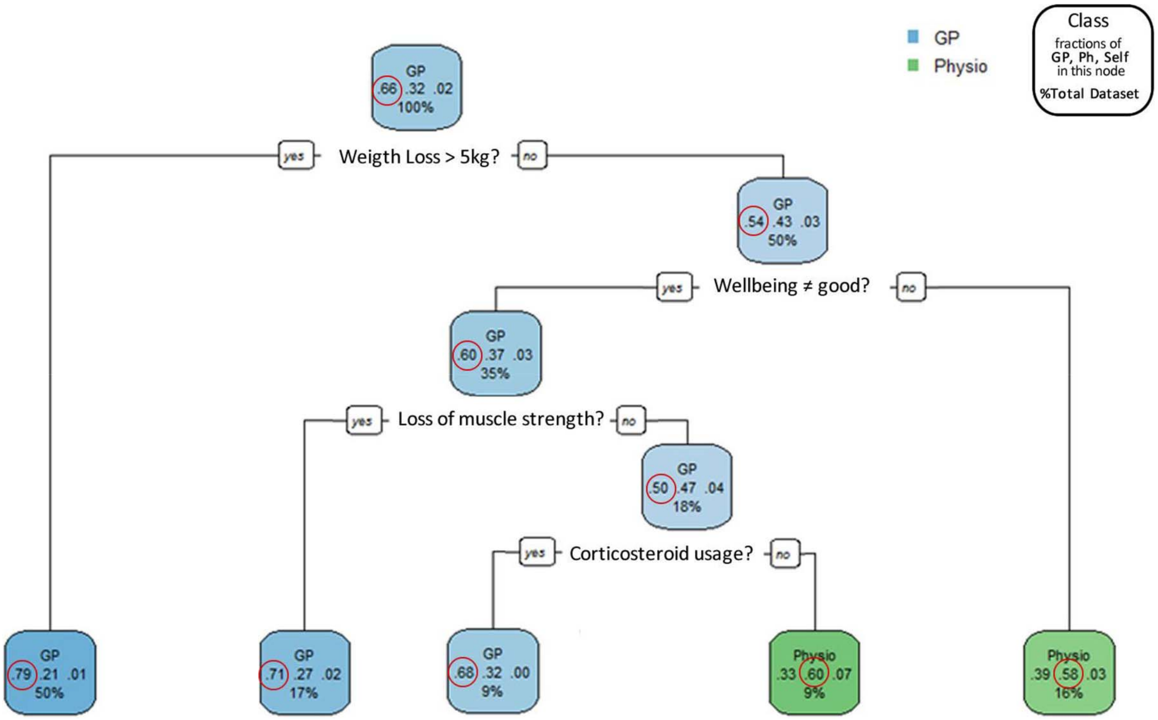


Fig. 4. Decision tree as generated in R on the training dataset. The class of a node/leaf in this tree is based on the highest fraction of a class in this node/leaf, which have been marked with a red circle in this figure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Performance of the decision tree during model validation and during model testing (Fig. 3), where G represents the class GP, P the class physiotherapy, S the class self-care, and C.I. is Confidence Interval.

		Decision tree		
		Calculated evaluation measures on the Validation dataset		
Prediction		**Reference**		
		G	**P**	**S**
	G	212	38	0
	P	74	62	0
	S	2	2	0
	Sensitivity	0.7361	0.6078	0.0000
	Specificity	0.6275	0.7431	0.9897
	Precision	0.8480	0.4559	0.0000
	Accuracy /95% C.I.	0.7026 /(0.6545, 0.7475)		

		Decision tree		
		Calculated evaluation measures on the Test dataset		
Prediction		**Reference**		
		G	**P**	**S**
	G	1	3	0
	P	4	26	0
	S	0	4	0
	Sensitivity	0.2000	0.7879	0.0000
	Specificity	0.9091	0.2000	0.8947
	Precision	0.2500	0.8667	0.0000
	Accuracy /95% C.I.	0.7105 /(0.5410, 0.8458)		

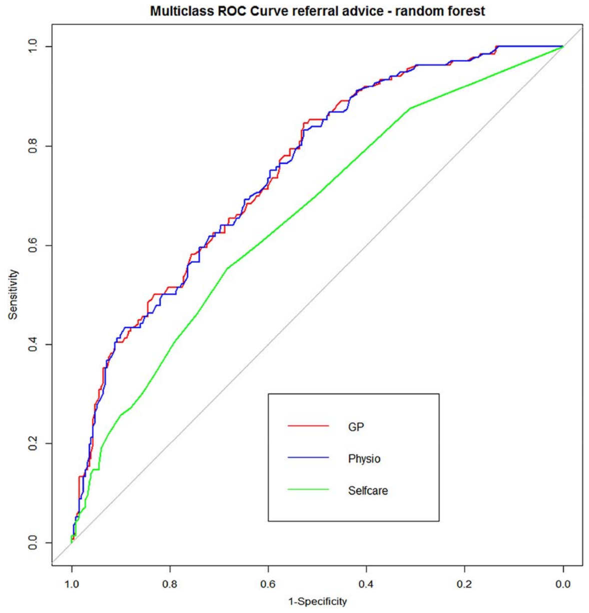


Fig. 5. The multiclass ROC Curve of the random forest.

referral advice better than just a random guess. When taking the majority referral class (GP advice) of the training dataset as default class, the baseline values of sensitivity during model validation and model testing are 0.65 and 0.11 respectively. This is because 65% of the cases advice in the training dataset, and 11% of the cases in the test dataset, were classified as GP advice. All measured sensitivities were higher than these baseline values. This means that all models learned some implicit knowledge of the provided referral advices in the training dataset.

Fig. 9 shows the estimated spread and mean of the accuracy, as well as of the kappa, for each model. The boosted tree appeared to be the best for predicting a referral advice with a fair accuracy (Kappa between 0.2 and 0.4). Next to this, Fig. 10 shows that the boosted tree performed best on accuracy both during model validation as well as during model testing (72% and 71% respectively). Furthermore, the averaged sensitivity and specificity of the boosted tree model were the highest during model testing, meaning that the boosted tree model performs best on a dataset with real-life cases.

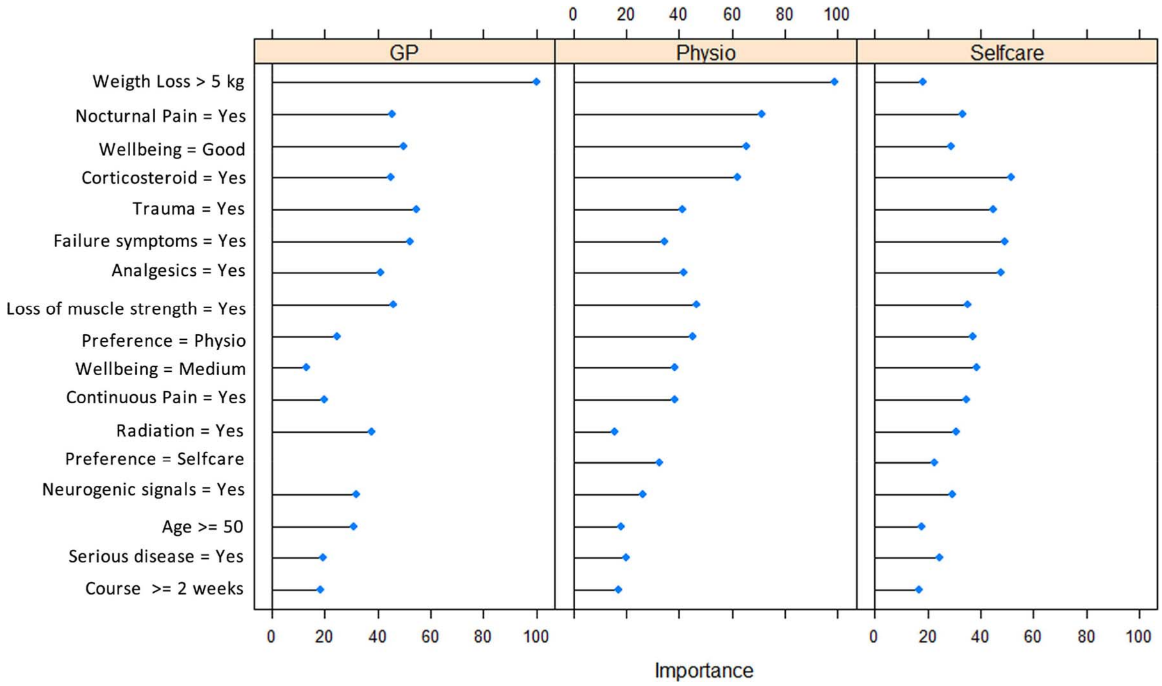


Fig. 6. Determined variable importance in the random forest for each class. The variable importance values are scaled to have a maximum value of 100.

Table 4
Performance of the random forest as estimated during model validation and during model testing (Fig. 3), where G represents the class GP, P the class physiotherapy, S the class self-care, and C.I. is Confidence Interval.

		Random Forest		
		Calculated evaluation measures on the Validation dataset		
Prediction		**Reference**		
		G	**P**	**S**
	G	240	10	0
	P	105	31	0
	S	2	2	0
	Sensitivity	0.6916	0.7209	0.0000
	Specificity	0.7674	0.6974	0.9897
	Precision	0.9600	0.2279	0.0000
	Accuracy /95% C.I.	0.6949 /(0.6465, 0.7402)		
		Random Forest		
		Calculated evaluation measures on the Test dataset		
Prediction		**Reference**		
		G	**P**	**S**
	G	3	1	0
	P	13	17	0
	S	4	0	0
	Sensitivity	0.1500	0.9444	0.0000
	Specificity	0.9444	0.3500	0.8947
	Precision	0.7500	0.5667	0.0000
	Accuracy /95% C.I.	0.5263 /(0.3582, 0.6902)		

4. Discussion

In this study, we explored the possibility of using machine learning in the design of a CDSS to support patients with a novel episode of LBP in their self-referral to primary care. At this moment, mainly patients with a higher education level, a shorter duration of symptoms and recurrent symptoms use the option of self-referral [16,17]. For a group of patients it is still unclear what to do first: consult a GP, consult a physiotherapist, or perform self-care first. It is important is to ensure that all patients receive the right interventions at the right moment to prevent that acute LBP becoming chronic [8] with considerable more impact for the patient and costs for the society [7].

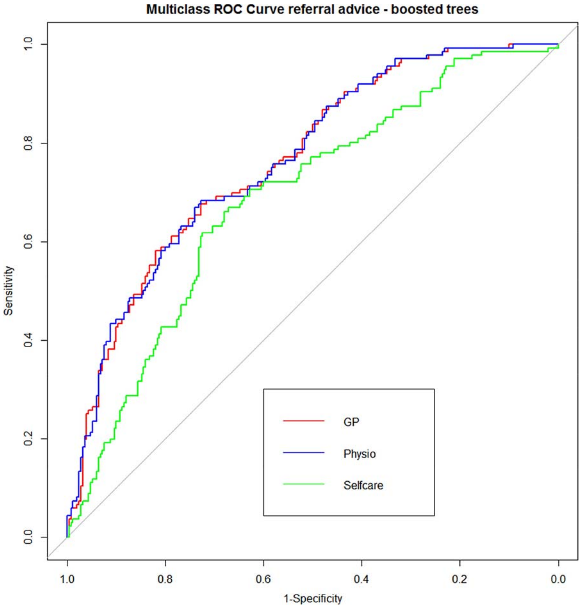


Fig. 7. The multiclass ROC Curve of the boosted tree.

A CDSS relays on computational models that can also be constructed and maintained based on machine learning [21]. This instead of a knowledge-based approach, in which a knowledge base and an inference engine are constructed and maintained based on knowledge elicited from literature and experts. This process of knowledge acquisition and maintenance can be very time consuming, and too expensive, and is also known as the "knowledge-acquisition bottleneck" [37]. When machine learning can be used in the design of our CDSS, we expect to avoid this kind of problems. Especially because digital data sources, as for example electronic health records, are becoming increasingly available. These sources contain data that can subsequently be used to train and maintain/improve the models.

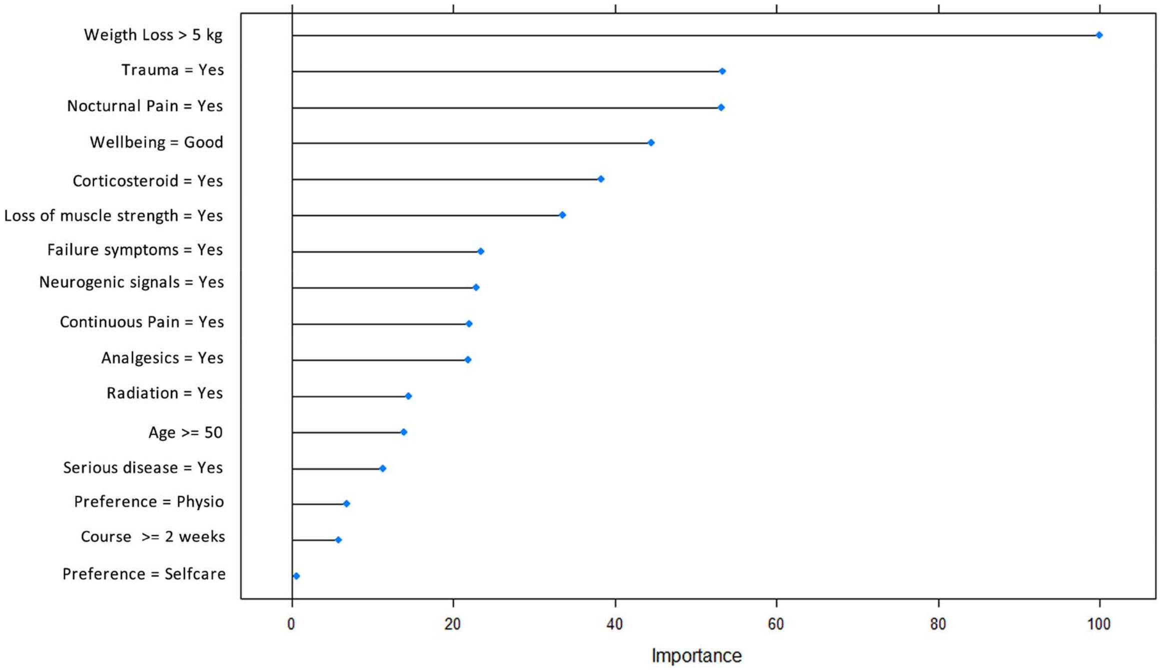


Fig. 8. Determined total variable importance in the boosted tree. The variable importance values are scaled to have a maximum value of 100.

Table 5
Performance of the boosted tree as estimated during model validation and during model testing (Fig. 3), where G represents the class GP, P the class physiotherapy, S the class self-care, and C.I. is Confidence Interval.

		Boosted tree		
		Calculated evaluation measures on the Validation dataset		
Prediction		**Reference**		
		G	**P**	**S**
	G	222	77	1
	P	28	59	3
	S	0	0	0
	Sensitivity	0.8880	0.4338	0.0000
	Specificity	0.4429	0.8780	1.0000
	Precision	0.7400	0.6556	NA
	Accuracy /95% C.I.	0.7205 / (0.6731, 0.7645)		
		Boosted tree		
		Calculated evaluation measures on the Test dataset		
Prediction		**Reference**		
		G	**P**	**S**
	G	1	4	3
	P	3	26	1
	S	0	0	0
	Sensitivity	0.2500	0.8667	0.0000
	Specificity	0.7941	0.5000	1.0000
	Precision	0.1250	0.8667	NA
	Accuracy /95% C.I.	0.7105 /(0.5410, 0.8458)		

During this study, we focused on the classification models decision tree, random forest and boosted tree. One should be aware that more types of classification models can be generated by machine learning algorithms. Other common machine learning algorithms are for example linear regression, neural networks, and support vector machines. Each machine learning algorithm has its own pros and cons [25,38] that may differ on the type of features used, (e.g., continuous, categorical). Decision tree is the machine learning algorithm that can handle both categorical and continuous features, and is used most for classification problems as decision trees are self-explanatory and easy to follow [28]. Therefore, we have chosen for decision tree, random forest and boosted tree − i.e. tree based models − for this study.

In our study, the performance measures of the three models were estimated twice: 1. during model validation with 30% of the training dataset, and 2. during model testing with a test dataset with real-life cases of LBP. The exploration with the models shows that the boosted tree performed best. The measured performances also show that all models provided a referral advice better than just a random guess, meaning that all models learned some implicit knowledge from the examples in the training dataset.

4.1. *Study limitations*

The distribution of the referral advice classes in the training dataset as well as in the test dataset was imbalanced. For the training dataset, the cases in the vignette study mainly contained serious factors indicating that the patient should see a GP [39]. Therefore, most cases in the training dataset were classified as "GP advice". Subsequently, the models in this study were trained with an overrepresentation of GP advices. In the test dataset most cases were classified as "physiotherapist advice". Despite the overrepresentation of the GP class in the training dataset, for the test dataset the sensitivities of the models still scored well on physiotherapy advice. Nevertheless, this wide variation in referrals (GP referral, physiotherapist referral and the very small number of self-care referral) will be an area to be improved in future work.

4.2. *Future research*

The study showed promising results on using machine learning in the design of our CDSS. However, before machine learning can really be used, we have to collect more cases classified as self-care to be sure that also the self-care advice can be predicted well. This is also the most interesting referral class, because there is an increasing interest in using digital interventions to support patient self-management in LBP [36]. When self-care can be predicted well, a next step is to provide patients with personalized information on how to cope with the LBP and what exercises may be helpful. In this, a web-based program for self-management of LBP can be deployed, just as for example the system that is

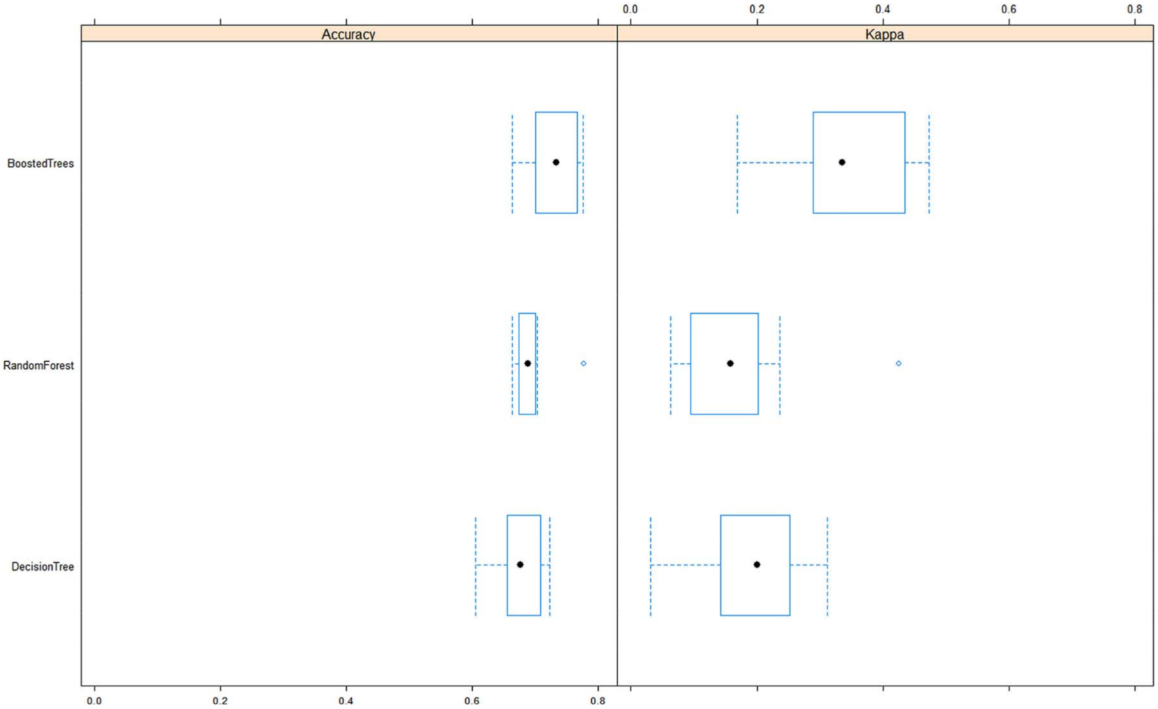


Fig. 9. Overview of the spread and the mean of the accuracy, as well as of the kappa, for each model.

used for patients with COPD [40].

5. Concluding remarks

Our study showed promising possibilities of using machine learning in the design of a CDSS to support patients with LBP in their self-referral process to primary care. CDSSs that support self-referral as well as further referral by healthcare professionals have the potential to decrease the current long waiting lines in healthcare in many countries. However, getting there is a long process and further study is needed on machine learning with larger data sets containing new cases, especially cases that are classified as self-care cases, to improve the model performances.

Funding

This work was conducted within the context of the eLabEL project, which aims to contribute to the sustainability of primary care by developing, implementing, and evaluating innovative, integrated telemedicine technology by means of a living lab approach. The eLabEL project is part of the Centre for Care Technology Research (www.caretechnologyresearch.nl). This work is partly funded by a grant from

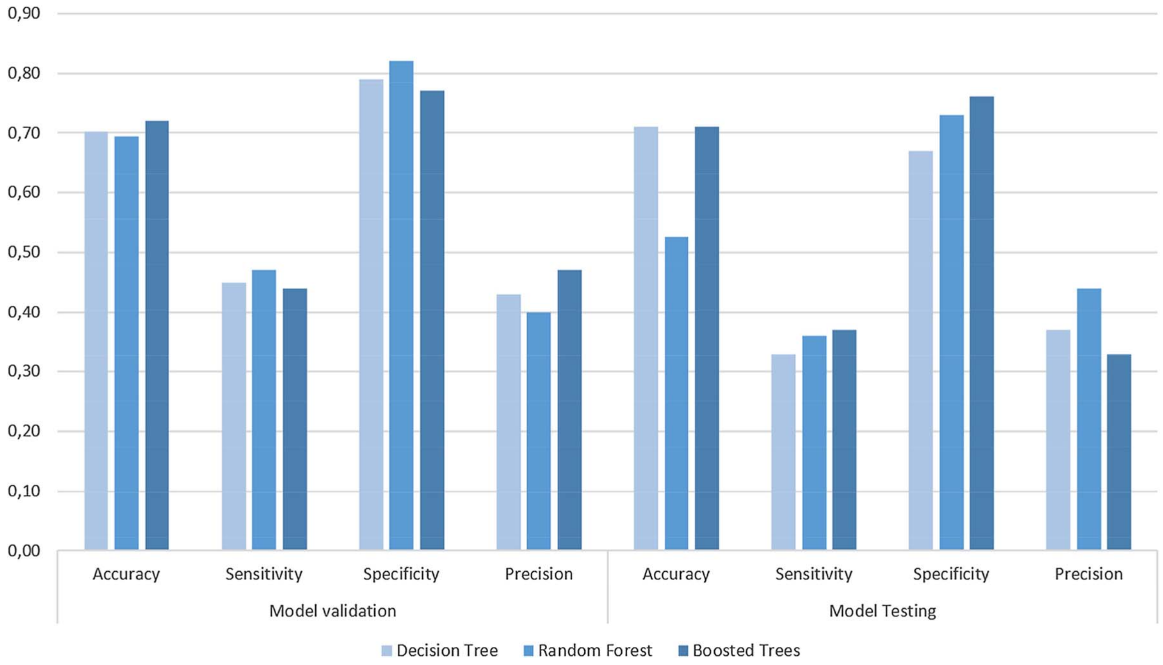


Fig. 10. Graphical presentation of the evaluation measures of the models as estimated during model validation as well as during model testing. The Accuracy is the total observed accuracy, and Sensitivity, Specificity, and Precision are the averaged Sensitivity, Specificity, and Precision of the three self-referral classes GP, Physio and Self-care.

the Netherlands Organisation for Health Research and Development (ZonMw), grant 10-10400-98-009.

Availability of data and materials

The used datasets are not publicly available in order to protect the anonymity of the participants. Datasets are available from the corresponding author on reasonable request.

Authors' contributions

All authors made substantial contributions to conception and design of this study. WH, LV, and HH discussed and designed the described research. LV and SR contributed to the acquisition of real-life cases of low back pain for the test dataset. WH trained the machine learning models, based on advices from MP and CG. WH drafted the initial manuscript. WH, LV, MP, RS, CG and HH participated in revising the manuscript critically for important intellectual content. All authors gave their final approval of the version to be submitted and any revised version.

Competing interests

WH, MP, CG and HH work for the University of Twente in Enschede, in the Netherlands. HH also works for Roessingh Research and Development, one of the participating companies in the eLabEL project. LV also works for the Roessingh Research and Development. RS works for the Groningen Spine Centre of the University Medical Centre Groningen in the Netherlands and the Saxion University of Applied Science in Enschede, also in the Netherlands. There are no financial or non-financial competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The study design to collect real-life case of low back pain was presented to the medical ethical committee (METC) Twente in Enschede, the Netherlands. We received a statement that ethics approval was not required for our study, as the normal healthcare path was not influenced and the patients remained anonymous to the researchers. All participants had given informed consent before filling in the online questionnaire.

Summary table

What was already known on the topic

Most people experience low back pain (LBP) at least once in their life and for some patients this evolves into a chronic condition. Because LBP causes considerable disability and financial burden globally, it is of importance to prevent the development of chronic LBP wherever possible. One way to prevent acute LBP from transiting into chronic LBP, is to ensure that patients receive the right interventions at the right moment starting with helping patients in their self-referral decision on what to do first: consult a GP, consult a physiotherapist, or perform self-care first. This is a classification process.

As self-referral can be seen as a classification process, supervised machine learning can be applied to design a classification model representing this process. This is supported by literature that shows that machine learning is increasingly used in healthcare informatics, also in the case of patient referral.

What this study added to this knowledge

The study showed promising possibilities of using machine learning to train models that can support patients with low back pain in their decision on self-referral to primary care.

Machine learning is a data-driven approach, and model performances highly depend on available training data. The currently found model performances can be improved further by increasing the training dataset with new cases, especially cases that are classified as self-care cases.

Acknowledgments

The authors would kindly like to thank all participants in the evaluation study - i.e. patients, doctors' assistants, general practitioners, and physiotherapists - for their time, effort, and willingness to cooperation in our research.

References

[1] G.E. Ehrlich, Low back pain, Bull. World Health Organ. 671 (2003) (81, 9).

[2] Damian Hoy, Christopher Bain, Gail Williams, Lyn March, Peter Brooks, Fiona Blyth, Anthony Woolf, Theo Vos, Rachelle Buchbinder, A systematic review of the global prevalence of low back pain, Arthr. Rheumat. 64 (6) (2012) 2028–2037.

[3] Laxmaiah Manchikanti, Vijay Singh, Frank J.E. Falco, Ramsin M. Benyamin, Joshua A. Hirsch, Epidemiology of low back pain in adults, Neuromodul.: Technol. Neural Interface 17 (S2) (2014) 3–10.

[4] B.W. Koes, M.W. Van Tulder, S. Thomas, Diagnosis and treatment of low back pain, BMJ: Br. Med. J. 332 (7555) (2006) 1430.

[5] Miranda L. van Hooff, Jan van Loon, Jacques van Limbeek, Marinus de Kleuver, The Nijmegen decision tool for chronic low back pain. Development of a clinical decision tool for secondary or tertiary spine care specialists, PLoS One 9 (8) (2014) e104226.

[6] Clermont E. Dionne, Kate M. Dunn, Peter R. Croft, Does back pain prevalence really decrease with increasing age? A systematic review, Age Ageing 35 (3) (2006) 229–234.

[7] Maria Vassilaki, Eric L. Hurwitz, Insights in public health: perspectives on pain in the low back and neck: global burden, epidemiology, and management, Hawai'i J. Med. Public Health 73 (4) (2014) 122.

[8] J.M. Fritz, J.D. Childs, R.S. Wainner, T.W. Flynn, Primary care referral of patients with low back pain to physical therapy: impact on future health care utilization and costs, Spine 37 (25) (2012) 2114–2121.

[9] Anthony Delitto, Richard E. Erhard, Richard W. Bowling, A treatment-based classification approach to low back syndrome: identifying and staging patients for conservative treatment, Phys. Ther. 75 (6) (1995) 470–485.

[10] B. Widerström, N. Olofsson, C. Boström, E. Rasmussen-Barr, Feasibility of the subgroup criteria included in the treatment-strategy-based (TREST) classification system (CS) for patients with non-specific low back pain (NSLBP), Man.Ther 23 (2016) 90–97.

[11] Bart W. Koes, Maurits van Tulder, Chung-Wei Christine Lin, Luciana G. Macedo, James McAuley, Chris Maher, An updated overview of clinical guidelines for the management of non-specific low back pain in primary care, Eur. Spine J. 19 (12) (2010) 2075–2094.

[12] J.C. Hill, D.G.T. Whitehurst, M. Lewis, S. Bryan, K.M. Dunn, N.E. Foster, K. Konstantinou, C.J. Main, E. Mason, S. Somerville, G. Sowden, K. Vohora, E.M. Hay, Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial, Lancet 378 (9802) (2011) 1560–1571.

[13] D.P. Gross, S. Armijo-Olivio, W.S. Shaw, K. Williams-Whitt, N.T. Shaw, J. Hartvigsen, Z. Qin, C. Ha, L.J. Woodhouse, I.A. Steenstra, Clinical decision support tools for selecting interventions for patients with disabling musculoskeletal disorders: a scoping review, J. Occup. Rehabil. 25 (December (4)) (2015) 675–782.

[14] T.J. Bury, E.K. Stokes, A global view of direct access and patient self-referral to physical therapy: implications for the profession, Phys. Ther. 93 (4) (2013) 449–459.

[15] I.C.S. Swinkels, M.K. Kooijman, P.M. Spreeuwenberg, D. Bossen, C.J. Leemrijse, C.E. van Dijk, R. Verheij, D.H. de Bakker, C. Veenhof, An overview of 5 years of patient self-referral for physical therapy in the Netherlands, Phys. Ther. 94 (12) (2014) 1985–1995.

[16] Jantine Scheele, Frank Vijfvinkel, Marijn Rigter, Ilse C.S. Swinkels, Sita M.A. Bierman-Zeinstra, Bart W. Koes, Pim A.J. Luijsterburg, Direct access to physical therapy for patients with low back pain in the Netherlands: prevalence and predictors, Phys. Ther. 94 (3) (2014) 363–370.

[17] N.E. Lankhorst, J.A. Barten, R. Meerhof, S.M.A. Bierma-Zeinstra, M. van Middelkoop, Characteristics of patients with knee and ankle symptoms accessing physiotherapy: self-referral vs general practitioner's referral, Physiotherapy (2017) Available online 24 May 2017 (in Press).

[18] C.A. Shaheed, B. McFarlane, C.G. Maher, K.A. Williams, J. Bergin, A. Matthews,

A.J. McLachlan, Investigating the primary care management of low back pain: a simulated patient study, J. Pain 17 (1) (2016) 27–35.

[19] W. Oude Nijeweme-d'Hollosy, L. Velsen, R. Soer, H. Hermens, Design of a web-based clinical decision support system for guiding patients with low back pain to the best next step in primary healthcare, Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) (2016) 229–239 (Vol. 5: HEALTHINF).

[20] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, third edition, Elsevier, 2012.

[21] Sumeet Dua, U. Rajendra Acharya, Prerna Dua (Eds.), Machine Learning in Healthcare Informatics, Springer, Berlin, 2014.

[22] S. Horng, D.A. Sontag, Y. Halpern, Y. Jernite, N.I. Shapiro, L.A. Nathanson, Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning, PLoS One 12 (4) (2017) p.e0174708.

[23] William Caicedo-Torres, Gisela García, Hernando Pinzón, A machine learning model for triage in lean pediatric emergency departments, Ibero-American Conference on Artificial Intelligence, Springer International Publishing, 2016, pp. 212–221.

[24] Z. Qin, S. Armijo-Olivo, L.J. Woodhouse, D.P. Gross, An investigation of the validity of the Work Assessment Triage Tool clinical decision support tool for selecting optimal rehabilitation interventions for workers with musculoskeletal injuries, Clin. Rehabil. 30 (3) (2016) 277–287.

[25] Sotiris B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised Machine Learning: A Review of Classification Techniques, (2007), pp. 3–24.

[26] Wendy Oude Nijeweme–d'Hollosy, Lex van Velsen, Karin G.M. Groothuis-Oudshoorn, Remko Soer, Hermie Hermens, Should I see a healthcare professional or can I perform self-care: self-referral decision support for patients with low back pain, Healthcare Informatics (ICHI), 2016 IEEE International Conference, IEEE, 2016, pp. 495–503.

[27] Christian Winther Topp, Søren Østergaard, Susan Søndergaard, Per Bech, The WHO-5 Well-Being Index: a systematic review of the literature, Psychother. Psychosom. 84 (3) (2015) 167–176.

[28] Lior Rokach, Oded Maimon, Data Mining with Decision Trees: Theory and Applications. World Scientific, (2014).

[29] G. Dietterich Thomas, Ensemble methods in machine learning, Multiple Classifier Syst. 1857 (2000) 1–15.

[30] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, Classification and Regression Trees, Wadsworth & Brooks, 1984.

[31] Leo Breiman, Random forests, Machine Learn. 45 (1) (2001) 5–32.

[32] Eric Bauer, Ron Kohavi, An empirical comparison of voting classification algorithms: bagging, boosting, and variants, Machine Learn. 36 (1) (1999) 105–139.

[33] https://www.r-project.org/, May 2017.

[34] https://www.rstudio.com/, May 2017.

[35] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.

[36] Barbara I. Nicholl, Louise F. Sandal, Mette J. Stochkendahl, Marianne McCallum, Nithya Suresh, Ottar Vasseljen, Jan Hartvigsen, et al., Digital support interventions for the self-management of low back pain: a systematic review, J. Med. Internet Res. 19 (5) (2017) e179.

[37] Eta S. Berner, Clinical Decision Support Systems 233 Springer Science+ Business Media, LLC, New York, 2007.

[38] R. Schapire, Machine Learning Algorithms for Classification, Princeton University, 2015, p. 10.

[39] J.B. Staal, E.J.M. Hendriks, M. Heijmans, H. Kiers, A.M. Lutgers-Boomsma, G. Rutten, M.W. van Tulder, J. Den Boer, R. Ostelo, J.W.H. Custers, KNGF-richtlijn Lage Rugpijn, Koninklijk Nederlands Genootschap Voor Fysiotherapie, De Gans Amersfoort, The Netherlands (Dutch), 2013.

[40] Monique Tabak, Marjolein Brusse-Keizer, Paul van der Valk, Hermie Hermens, Miriam Vollenbroek-Hutten, A telehealth program for self-management of COPD exacerbations and promotion of an active lifestyle: a pilot randomized controlled trial, Int. J. Chron. Obstruct. Pulmon. Dis. 9 (2014) 935.