Journal of Clinical Epidemiology 177 (2025) 111593

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Prediction models for outcomes in people with low back pain receiving conservative treatment: a systematic review

Rubens Vidal[a], Margreth Grotle[b,c], Marianne Bakke Johnsen[b], Louis Yvernay[b], Jan Hartvigsen[d,e], Raymond Ostelo[f,g], Lise Grethe Kjønø[b], Christian Lindtveit Enstad[b], Rikke Munk Killingmo[b], Einar Henjum Halsnes[b], Guilherme H.D. Grande[a], Crystian B. Oliveira[a,*]

[a]*Faculty of Medicine, University of West Paulista (UNOESTE), Presidente Prudente, Brazil*
[b]*Faculty of Health Sciences, Department of Rehabilitation Science and Health Technology, Oslo Metropolitan University, Oslo, Norway*
[c]*Division of Clinical Neuroscience, Department of Research, Innovation and Education, Oslo University Hospital, Oslo, Norway*
[d]*Center for Muscle and Joint Health, Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark*
[e]*Chiropractic Knowledge Hub, University of Southern Denmark, Odense, Denmark*
[f]*Department of Health Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*
[g]*Department of Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit & Amsterdam Movement Sciences, Amsterdam, The Netherlands*

Accepted 3 November 2024; Published online 9 November 2024

Abstract

Objectives: To identify, critically appraise and evaluate the performance measures of the available prediction models for outcomes in people with low back pain (LBP) receiving conservative treatment.

Study Design and Setting: In this systematic review, literature searches were conducted in Embase, Medline, and cumulative index of nursing and allied health literature from their inception until February 2024. Studies containing follow-up assessment (eg, prospective cohort studies, registry-based studies) investigating prediction models of outcomes (eg, pain intensity and disability) for people with LBP receiving conservative treatment were included. Two independent reviewers performed the study selection, the data extraction using the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies, and risk of bias assessment using the Prediction model Risk of Bias Assessment. Findings of individual studies were reported narratively taking into account the discrimination and calibration measures of the prediction models.

Results: Seventy-five studies developing or investigating the validity of 216 models were included in this review. Most prediction models investigated people receiving physiotherapy treatment and most models included sociodemographic variables, clinical features, and self-reported measures as predictors. The discriminatory capacity of the internal validity of the 27 prediction models for pain intensity varied greatly showing a c-statistic ranging from 0.48 to 0.94. Similarly, the discriminatory capacity for 31 models for disability had the same pattern showing a c-statistic ranging from 0.48 to 0.86. The calibration measures of the internal validity of the prediction models predicting pain intensity and disability showed to be adequate. Only one of 3 studies testing the external validity of models to predict pain intensity and disability and reported both discrimination and calibration measures, which showed to be inadequate. The prediction models predicting the secondary outcomes (eg, self-reported recovery, quality of life, return to work) showed varied performance measures for internal validity, and only 2 studies tested the external validity of models although they did not provide performance the performance measures.

Conclusion: Several prediction models have been developed for people with LBP receiving conservative treatment; however, most show inadequate discriminatory validity. A few studies externally validated the prediction models and future studies should focus on testing this before implementing in clinical practice.

Keywords: Low back pain; Prediction models; Prognosis research; Conservative treatment; Physiotherapy; Pain management

Funding: This study is part of the AID-Spine project funded by the Norwegian Research Council (Grant number: 324,915)

* Corresponding author. University of Western São Paulo (UNOESTE), Rua José Bongiovani, 700, Presidente Prudente, São Paulo CEP 19050-920, Brazil.

https://doi.org/10.1016/j.jclinepi.2024.111593

What is new?

Key findings

- Several prognostic models have been developed for people with low back disorders receiving conservative treatment.
- Internal validity showed a wide variation in calibration and discriminatory capacity, although some models showed adequate performance measures to predict pain and disability as well as the secondary outcomes.
- Few studies external validated the prognostic models and showed inadequate performance measures.
- All studies were assessed as having high risk of bias.
- High-quality studies focusing on externally validating the existent models are warranted.

What this adds to what is known?

- The included studies had some methodological flaws including the selection of predictors using univariate analysis and lack of discriminatory and calibration measures.
- Several models were developed and had their internal validity tested, and some showed adequate discriminatory and calibration measures.
- Although only a few studies aimed to externally validate prognostic models, none showed adequate calibration and discrimination measures to move to the next stage and be tested in clinical practice.

What is the implication and what should change now?

- Future studies aiming to develop prognostic models should improve the methodology for selecting variables and providing adequate information on the performance of the models.
- Some models showed to have acceptable discriminatory and calibration measures, so further studies should focus on testing their external validity instead of developing new models.
- The use of prediction models for decision making of people with low back pain receiving conservative treatment is still not recommended.

1. Introduction

Prediction models have been proposed as an important approach for identifying people for a specific treatment in different populations. From a set of candidate predictors, they are developed using multivariable statistical approaches and more recently, machine learning methods. These models are validated in the same sample to assess the internal validation and then using an external sample to assess the external validation [1]. The final stage is to test the clinical impact of the implementation of the prediction model in routine clinical practice [1]. One example of a prognostic model is the body mass index, airflow obstruction, dyspnea, and exercise capacity index, developed to predict mortality in people with chronic obstructive pulmonary disease [2]. This model has been externally validated by 13 studies [3]. This approach could also help identifying people with low back pain (LBP) experiencing poor or successful outcomes after receiving specific treatments.

The use of interventions with no or little benefits, when the harms outweigh the benefits, or their added costs do not result in additional benefits is considered low-value care [4], and this situation is prevalent across musculoskeletal conditions, especially in people with LBP. However, patient's response after treatment may differ considering specific characteristics and only those experiencing poor treatment outcomes should be considered as receiving low-value care. The reason is that these people would probably look for additional treatments and, consequently, increase the costs related to LBP. In fact, the amount of low-value care given to this population can be responsible to the enormous economic impact caused by LBP on health-care systems [5]. Previous studies reported that direct costs related to LBP were USD $134.5 billion in 2016 in the United States [6], USD $71,4 million from 2012 to 2016 in Brazil [7], and €740 million in 2011 in Sweden [8]. To provide better treatment options, we need systems that could help us identify these patients experiencing poor treatment outcomes.

Several prediction models have been developed in the last years to predict outcomes of people with LBP. Although previous reviews investigated the evidence around the use of prognostic models for people with LBP [9−11], the searches were conducted 10 years ago [9], focusing on people who did not receive any treatment [11], or on those receiving surgical treatment [10]. However, the available prediction models for people with LBP receiving conservative treatments are unknown. For this review purposes, we defined conservative treatment as any nonsurgical approach, including pharmacological interventions, spinal injections, physical therapy, exercise, manual

therapy, patient education/advice and others. Therefore, the aim of this review was to identify and critically appraise the available prediction models to predict successful/poor health outcomes in people with LBP receiving conservative treatment. The critical appraisal included an assessment of the performance measures (discrimination and calibration measures) for those studies testing the internal or external validity of the prediction models.

2. Methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses recommendations to report this review [12]. This review was prospectively registered in PROSPERO before starting to conduct it (number: CRD42022370503).

2.1. Searches

Literature searches were conducted in the following electronic databases from their inception until February/2024: EMBASE (via Embase.com), MEDLINE (via Ovid), and cumulative index of nursing and allied health literature (via EBSCO). We used a combination of terms related to LBP and lower back disorders (eg, LBP, sciatica, scoliosis, and others), prediction models (eg, prognosis, prediction) and spinal conservative treatments. We modified the search filters available in the literature [13] and included 2 terms (ie, predict*, prognos*) to identify relevant prediction modeling studies in the field. We also screened the references of relevant publications in the field to identify any additional potentially eligible studies. There were no restrictions regarding language and date of publications. Appendix 1 details the search strategy performed in electronic databases.

2.2. Eligibility criteria

We included studies developing (ie, selection of a set of predictors using a multivariable approach regardless of statistical approach) and/or validating (ie, the investigation of the validity of the model in the same or different sample population than the model was developed) prediction models predicting outcomes of people with LBP receiving specific conservative treatments. All longitudinal study designs conducting patient follow-up, such as prospective cohort studies, randomized controlled trials, or registry-based studies, were eligible. We adhered to the definition of prediction model research as used in the PROGRESS framework, in which prediction models aim to identify the probability of a specified outcome (endpoint) occurring in a specific population, using a multivariable approach including a set of predictors measured at baseline [1]. Studies investigating risk stratification tools (eg, STarT back screening tool, Örebro Musculoskeletal Pain Questionnaire) were considered eligible. However, we excluded studies investigating individual prognostic factors [14], studies aiming to identify people with LBP, and studies recruiting people seeking primary care without specifying the conservative treatment.

Studies recruiting adults (aged $\geq$18 years) with LBP reporting back pain and/or radiating leg pain receiving conservative treatments (eg, physiotherapy, exercise, injection, medication) were considered eligible. LBP was defined as people having herniated discs, spinal stenosis, scoliosis, spondylosis, spondyloarthrosis, and other degenerative disc disorders, or nonspecific LBP. We did not include studies investigating LBP related to a serious spinal pathology (eg, infection, tumor, fracture).

The primary outcomes of this review were successful or poor health outcome based on pain intensity (measured using, eg, Numerical Rating Scale, visual analog scale, McGill Pain Questionnaire, and others), or disability (measured using, eg, Roland Morris Disability Questionnaire, Oswestry Disability Index, Quebec back pain disability scale, and others). A successful health outcome was defined as reaching clinical improvement based on a predefined cut-off of pain and/or disability (eg, more than 30% of improvement from baseline, or a score lower than 2 out of 10 points at follow-up). In contrast, poor health outcome was defined as the lack of clinical improvement considering a cut-off point based on measures of pain and disability (eg, less than 30% of improvement from baseline, or a follow-up absolute score greater than a prespecified cut-off estimate). Anticipating that definitions of successful or poor outcome vary greatly in the literature, we also accepted the studies' definition based on mean improvement or an anchor-based method. As secondary outcomes, we included studies predicting other outcomes, such as self-reported recovery, quality of life, and others.

2.3. Data extraction

Data extraction was performed considering the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies [15]. Two independent authors extracted the following information from the included studies: characteristics of the study (eg, country, source, study design), characteristics of the participants (sample size, age and gender); treatment (type, length, frequency, dose), and outcomes (instrument, definition, time point follow-up assessment), predictors, missing values, model development, performance of the model, model validation, and interpretation of the model. We extracted discrimination (the ability of distinguishing with a successful or poor health outcome) and calibration measures (the extent that the expected outcomes agree with the observed outcomes) of studies assessing the performance measures of the internal and external validation of prediction models [16]. For discrimination, we preferentially extracted the concordance C-statistics or area under the curve from the receiving operating characteristic curve, when available.

For calibration, we extracted the calibration intercept, calibration slope, and the *P* value of the Hosmer-Lemeshow test, when available. If insufficient data were provided, additional information were requested to the corresponding authors or estimates were made using pre-established methods [16].

2.4. Risk of bias assessment

Prediction model Risk of Bias Assessment (PROBAST) was used to assess the methodological quality (risk of bias) and relevance (applicability) of the included studies investigating prognostic models across the 4 domains [17]: participants, predictors, outcome and analysis. The scale contains 20 items which can be rated as yes, probably yes, no, probably no, no information. Each domain is rated as low, high, or unclear risk of bias. In pairs, 2 independent authors performed the assessment, and, in case of disagreement, a third reviewer was consulted to arbitrate the discussion and resolve it. A study was judged as having high risk of bias when at least 1 of the domains was judged as having high risk of bias.

2.5. Data synthesis

Considering that included studies developed or validated individual models for a great variety of interventions and predicted different outcomes, the performance measures of the models were narratively described. We described the models considering the type of intervention and outcomes predicted. However, we generated a forest plot displaying the c-statistics and 95% CIs for included studies investigating the internal validity of prediction models. Calibration intercept and slope measures close to 1 and 0 represent correct calibration, respectively [18]. The graph was created using the RStudio (version 1.2.5042) through R software (version 4.0.2).

Although there are some adapted versions of the grading of recommendations assessment, development and evaluation approach for prognostic factors, the guidance to assess the certainty of evidence for systematic reviews of prognostic models using the grading of recommendations assessment, development and evaluation approach is still incomplete. Therefore, we did not assess the certainty of evidence in this review.

3. Results

Electronic searches performed in February 2024 retrieved 15,232 records after removing duplicates. After screening of titles and abstracts, we assessed the full text of 172 potentially eligible studies. The reasons for exclusion after full-text assessment were because studies investigated individual prognostic factors (n = 16), did not develop a prediction model (n = 57), did not perform a follow-up assessment (n = 8), did not include exclusively people with LBP (n = 13), or did not specifically investigate conservative treatments (n = 3). Appendix 2 details the reason for exclusion for each full text assessed. Finally, 75 studies were included in this review. Figure 1 details the processes of this review.

Table 1 summarizes the main characteristics of the studies. Most studies (n = 34, 45%) were conducted in Europe. The median sample size was 238 ranging from 30 to 154,167 participants, and the mean age ranged from 18 to 88. Most studies developed/validated prediction models for patients receiving spinal manipulation (n = 16 studies), exercise (n = 16 studies) or physiotherapy without specifying the treatments' components (n = 13 studies). The most common outcomes investigated by included studies were disability (n = 38 studies, 51%), pain intensity (n = 30 studies, 40%), and return to work (n = 13 studies, 17%). The time point follow-up assessment ranged from 9 days to 24 months. Twenty-four studies (32%) focused on developing prediction models, 48 studies (65%) on developing and internally validating prediction models, and 4 studies (5%) on developing and externally validating prediction models. Appendix 3 details the characteristics of the included studies.

Appendix 4 details the risk of bias of included studies using the PROBAST. The overall risk of bias of all included studies were assessed as having a high risk of bias, mostly because they failed to meet the domain related to analysis by selecting the predictors to be included in the model using a univariate approach (49 out of 75, 65%) or failed to report both calibration and discrimination measures (47 out of 75, 63%). In addition, more than half of included studies (40 out of 75, 53%) also did not handle the missing data appropriately because they included only complete cases or did not use an appropriate method to handle it. Regarding other domains, nearly all studies met the items related to the domains of participants, predictors, and outcome domains. Most studies (52 out of 77, 67%) were judged as having low risk of bias on applicability.

3.1. Development

Twenty-four studies reported on development of 82 models. Most were prospective cohort studies (n = 14) conducted in Europe (n = 12) and North America (n = 7). The prediction models were developed using multivariable regression analysis including a wide range of predictors. Most studies developed prediction models containing a set of predictors including sociodemographic variables (eg, age, gender, education levels), clinical features (eg, duration of symptoms, number of previous episodes), and self-reported measures (e.g., Oswestry

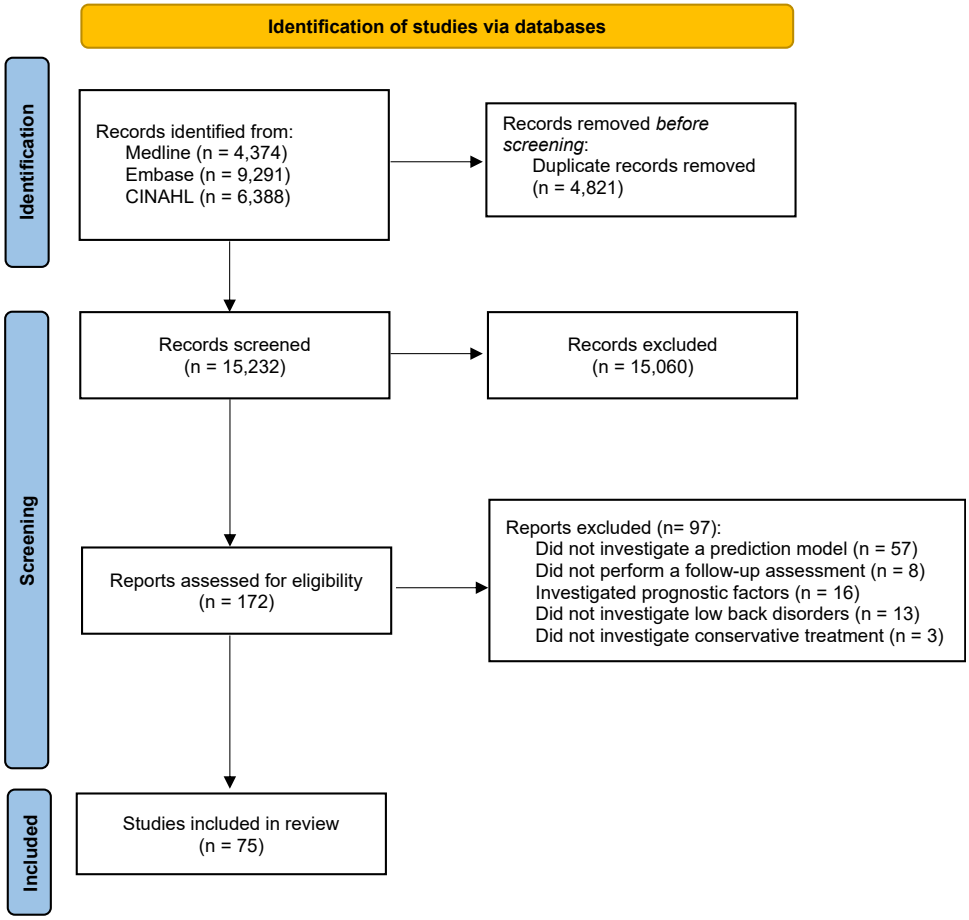

Figure 1. Flowchart of review's processes.

Disability Index, Roland Morris Disability Questionnaire, Numerical Pain Rating Scale, and Fear-Avoidance Behaviour Questionnaire). A few studies also included a measure after baseline assessment (eg, pain intensity after 1 week).

3.2. Primary outcomes

3.2.1. Internal validation

Forty-eight studies investigated the internal validity of 128 prediction models. Most studies focused on predicting pain intensity (n = 17 out of 75, 22%) or disability levels (n = 19 out of 75, 25%) for patients with LBP receiving spinal manipulation (n = 12 out of 75, 16%), exercises (n = 10 out of 75, 13%) and physiotherapy (n = 9 out of 75, 12%). The discrimination measures were mainly reported by included studies using the c-statistics. Figures 2 and 3 detail the discrimination measures reported by models predicting the primary outcomes pain intensity and disability, respectively. The discriminatory capacity varied greatly among the 27 predictions models for pain intensity with the c-statistic ranging from 0.48 to 0.94. The best predictions models of pain intensity with a c-statistic greater than 0.80 investigated people receiving multidisciplinary treatment or physiotherapy [19,24,25]. Similarly, the discriminatory capacity of the 31 prediction models for disability also varied greatly, as the c-statistic of the models ranged from 0.48 to 0.86. The best prediction models of disability with a c-statistic greater than 0.80 investigated people receiving physiotherapy or exercise.

3.3. External validation

Three studies [26–28] investigated the external validation of five models predicting pain (n = 2 out of 5 models, 40%) and disability (n = 3 out of 5 models, 60%), for people receiving spinal manipulation *plus* exercise, physiotherapy or pharmacological treatment, and epidural spine injections.

Traeger et al (2016) [28] was the only study reporting the discrimination and calibration measures of the prediction of pain and disability of people receiving physiotherapy or pharmacological treatment (paracetamol 4 g/day). The calibration measures showed to be inadequate for predicting low pain intensity (calibration intercept: 0.89; calibration slope: −0.59), high pain intensity (calibration intercept: 0.74; calibration slope: −0.81), and disability (calibration intercept: 0.99; calibration slope: −0.86). The discriminatory capacity

Table 1. Characteristics of the studies included in the systematic review

Characteristics	Models development	Internal validation	External validation
Number of studies	24 studies	48 studies[a]	4 studies[a]
Number of models	82 models	128 models	6 models
Period of publication	1998 to 2022	2002 to 2023	2004 to 2013
Study design			
Prospective Studies	14 studies	34 studies	1 study
Randomized Clinical Trial	10 studies	7 studies	2 studies
Registry-Based Studies	-	7 studies	1 study
Source of data			
Primary Care	3 studies	7 studies	1 study
Secondary Care	9 studies	19 studies	2 studies
Tertiary Care	4 studies	5 studies	1 study
Not mentioned	-	3 studies	-
Others	8 studies	14 studies	-
Sample size, median [range]	218.5 [53 to 934]	349 [30 to 154,167]	196.5 [131 to 1652]
Outcomes[b]			
Pain	12 studies	17 studies	1 study
Disability	16 studies	19 studies	3 studies
Work-related outcomes	1 study	11 studies	1 study
Quality of life	2 studies	5 studies	-
Global improvement	4 studies	11 studies	-
Others	4 studies	2 studies	-
Country			
Europe	12 studies	22 studies	-
North America	7 studies	13 studies	3 studies
Asia and Oceania	2 studies	13 studies	1 study
South America	3 studies	-	-
Type of intervention			
Physiotherapy[c]	4 studies	9 studies	-
Exercises	6 studies	10 studies	-
Spinal manipulation	4 studies	12 studies	-
Cognitive Behavioural Program	3 studies	4 studies	-
Medicines	1 study	6 studies	-
Multidisciplinary programs	2 studies	4 studies	1 study
Combined Therapies	3 studies	3 studies	2 studies
Others	1 study	-	1 study

[a] One study investigated the internal and external validity in the same study.
[b] The number of studies does not equal to the total number studies because some studies developed or validated predictions models for more than one outcome in the same study.
[c] Prediction models on physiotherapy which did not specify a specific treatment modality (eg, exercise).

of the model showed c-statistics lower than 0.7 for the outcomes. The other 2 studies [26,27] did not report quantitative data related to discrimination or calibration measures.

3.4. Secondary outcomes

3.4.1. Internal validity

Eight studies [24,29–35] reported discrimination measures for 12 prediction models for work-related outcomes. The c-statistic of the prediction models of work-related outcomes ranged from 0.51 to 0.85. Three models had adequate calibration measures for three models (calibration slope from 0.88 to 0.91) [30,31]. Seven models showed to have a *P* value higher than 0.05 in the Hosmer-Lemeshow test.

Seven studies [36–42] reported discrimination measures for 20 prediction models for global improvement. The c-statistic of the prediction models of global improvement ranged from 0.55 to 0.76. None reported calibration measures.

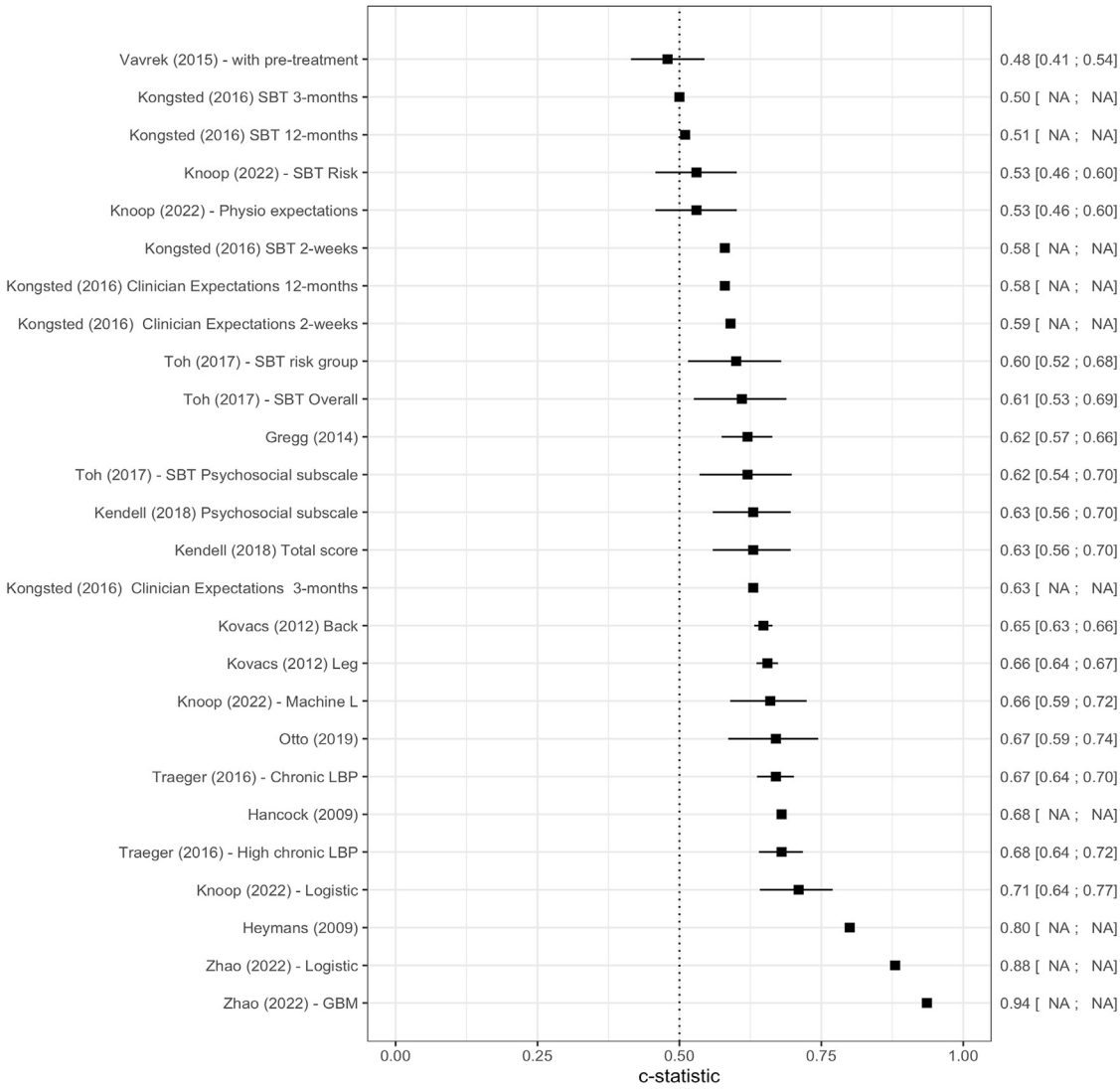

Figure 2. Area under curve (or c-statistics) of the included studies investigating pain intensity as an outcome-measure.

Three studies [41,43,44] reported discrimination measures for 5 prediction models for quality of life. The c-statistic of the prediction models of work-related outcomes ranged from 0.73 to 0.88. One study [43] reported the calibration plots showing a calibration curve close to the ideal line.

The remaining studies did not provide any discrimination or calibration measures or had an inadequate discriminatory capacity.

3.5. External validity

One study [45] investigated the external validation of a prediction model to return to work, but they did not provide the discrimination or calibration measures of the models.

4. Discussion

Our results indicate that several prediction models have been developed to predict clinical outcomes in patients with LBP undergoing conservative treatment such as physiotherapy treatment, spinal manipulation and exercise. All studies were assessed as having high risk of bias, indicating that the evidence around this topic should be interpreted with caution. Regarding internal validity, there was great variation in calibration and discriminatory capacity, although some models showed adequate performance measures to predict pain and disability. For external validity, only 6 models were tested and they either they did not have adequate performance measures or did not report quantitative data to be evaluated.

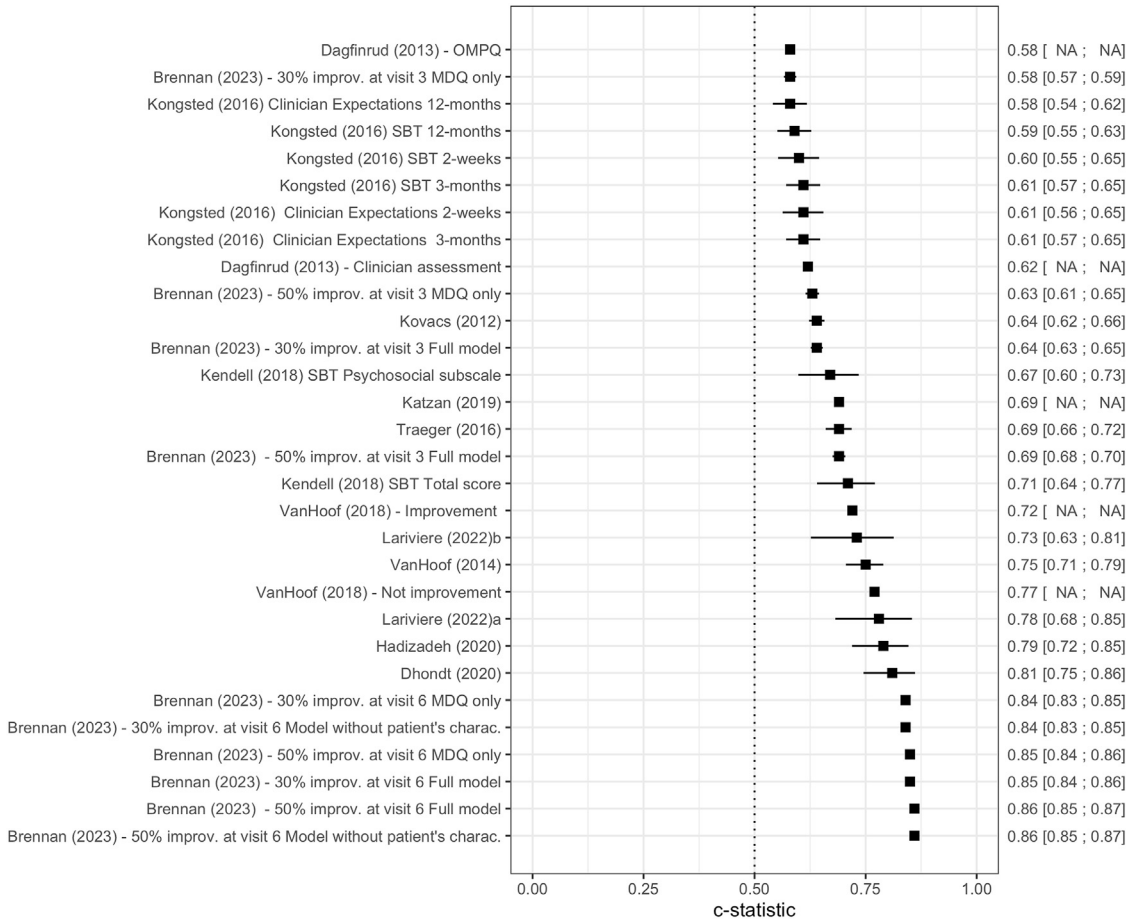

Figure 3. Area under curve (or c-statistics) of the included studies investigating disability as an outcome measure. Regarding calibration measures, 5 studies [19–23] investigating nine models reported a *P* value higher than 0.05 in the Hosmer-Lemeshow test, which indicates that there were no differences between the observed and expected event rates.

Previous reviews have been published investigating prediction models in people with LBP. McIntosh (2018) found that no prediction models for LBP (21 studies) were internally or externally validated [9]. In contrast, our review found that 63% of the included studies internally validated their models and 5% investigated the external validation. Another systematic review [11] revealed that some prediction models investigating patients with acute LBP did show adequate discrimination and only a few studies reported the calibration measures. More recently, a review [46] focused on prediction models for people with chronic LBP seeking primary care and the authors found similar methodological flaws (eg, handling missing data, selection of predictors) and a lack of studies reporting the calibration measures. Therefore, our results align with the findings of previous reviews indicating that an increasing number of prediction models have been validated, although most did not appropriately report the performance measures.

Some models showed good performance measures for prediction of pain intensity and disability of people with LBP receiving conservative treatment. Zhao (2022) [47] developed 2 models using different statistical methods (ie, logistic regression and another using gradient boosting, a machine learning–based technique using gradient boosting) and showed the best discrimination measures for predicting pain intensity and disability. Both consisted of 14 predictors that included age, gender, and imaging characteristics obtained through nuclear magnetic resonance and predicted pain intensity after the ninth physiotherapy session [47]. However, did not report the calibration measures of these models. In contrast, Lariviere (2022) [21] investigated 2 clinical prediction rules for disability in patients submitted to lumbar stabilization exercises. The models showed an appropriate discriminatory capacity and no difference between the observed and expected event rates. Nevertheless, we believe that these models could be prioritized in future externally validated efforts and then tested in clinical practice.

The strengths of this review include the prospective registration and comprehensive search strategy, out of which we were able to identify 75 included studies. Another strength was the descriptive analysis of the available prediction models considering the type of conservative treatment and outcome, which provided a broad perspective of the evidence in this field. This review has also some limitations. Firstly, we could

not exclude the possibility of missing studies as the topic is largely broad and we did not conduct searches in other sources (eg, gray literature). However, this could be considered a limitation of any systematic review. In addition, we identified a large number of included studies which would be representative of the existent studies in this area. Another limitation is the use of the PROBAST tool to assess the risk of bias of the included studies, which was only proposed recently. While this may partially explain why all included studies were classified as high risk of bias, authors must propose their studies using guidelines for conducting prediction studies (eg, PROGRESS framework) [1]. Lastly, although we did not make the data and codes available in a public repository, they are available upon reasonable request.

There is wide room for improvement in the area of prediction models after conservative treatment for LBP. First, future studies should address the most common methodological flaws identified in the included studies. This includes providing sufficient information regarding the performance measures, information on how missing data were handled, and using pre-existing knowledge of the possible candidates and their association with the outcome [18]. Considering the type of intervention used by the investigators in the included studies, there was insufficient information on some intervention characteristics, such as frequency, duration, type, intensity and dosage. This can also be considered a limitation, as it may affect the future efforts of externally validating the prediction models. Future studies should also provide a detailed description of the interventions using previously published checklists, such as the TIDieR checklist [48].

Future studies should also focus on externally validating the existent promising prediction models, to potentially refer patients to the most appropriate care. The models showing a good discrimination capacity for predicting pain intensity and disability investigated combined therapies as the conservative treatment (eg, physiotherapy associated with medication). This scenario would be a more realistic perspective of what the patient would experience in daily practice. In addition, the best models focused on predicting clinical outcomes at short-term follow-up (ie, 3 months after baseline assessment). These models also included a set of predictors combining demographic data (eg, age, gender), clinical characteristics (eg, pain intensity, radiated pain, emotional factors) and image exams (eg, magnetic resonance imaging). These predictors can be obtained in clinical practice, which can help overcome the barriers raised by clinicians about using prognostic models in clinical practice [49].

5. Conclusion

Many prediction models have been developed to predict clinical outcomes in patients with LBP undergoing conservative treatment. Regarding internal validity, there was a wide variation in calibration and discriminatory capacity, although a few models showed to have adequate performance measures. For external validity, only 6 models were tested and either they did not have adequate performance measures, or they did not report quantitative data to be evaluated. High-quality studies focusing on externally validating the existent models are warranted.

CRediT authorship contribution statement

Rubens Vidal: Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Formal analysis, Data curation. **Margreth Grotle:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marianne Bakke Johnsen:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Louis Yvernay:** Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation. **Jan Hartvigsen:** Writing – review & editing, Resources, Methodology, Conceptualization. **Raymond Ostelo:** Writing – review & editing, Methodology, Formal analysis, Data curation. **Lise Grethe Kjønø:** Writing – review & editing, Resources, Methodology, Formal analysis, Data curation. **Christian Lindtveit Enstad:** Writing – review & editing, Resources, Methodology, Formal analysis, Data curation. **Rikke Munk Killingmo:** Writing – review & editing, Methodology, Formal analysis, Data curation. **Einar Henjum Halsnes:** Writing – review & editing, Methodology, Formal analysis, Data curation. **Guilherme H.D. Grande:** Writing – review & editing, Methodology, Formal analysis, Data curation. **Crystian B. Oliveira:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

There are no competing interests for any author.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111593.

Data availability

Data will be made available on request.

References

[1] Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS)

3: prognostic model research. PLoS Med 2013;10(2):e1001381. https://doi.org/10.1371/journal.pmed.1001381.

[2] Celli BR, Cote CG, Marin JM, Casanova C, Montes de Oca M, Mendez RA, et al. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. N Engl J Med 2004;350(10):1005−12. https://doi.org/10.1056/NEJMoa021322.

[3] Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. BMJ 2019;367:l5358. https://doi.org/10.1136/bmj.l5358.

[4] Elshaug AG, Rosenthal MB, Lavis JN, Brownlee S, Schmidt H, Nagpal S, et al. Levers for addressing medical underuse and overuse: achieving high-value health care. Lancet 2017;390(10090):191−202. https://doi.org/10.1016/S0140-6736(16)32586-7.

[5] Foster NE, Anema JR, Cherkin D, Chou R, Cohen SP, Gross DP, et al. Prevention and treatment of low back pain: evidence, challenges, and promising directions. Lancet 2018;391(10137):2368−83. https://doi.org/10.1016/s0140-6736(18)30489-6.

[6] Dieleman JL, Cao J, Chapin A, Chen C, Li Z, Liu A, et al. US health care spending by payer and health condition, 1996-2016. JAMA 2020;323(9):863−84. https://doi.org/10.1001/jama.2020.0734.

[7] Carregaro RL, da Silva EN, van Tulder M. Direct healthcare costs of spinal disorders in Brazil. Int J Public Health 2019;64(6):965−74. https://doi.org/10.1007/s00038-019-01211-6.

[8] Olafsson G, Jonsson E, Fritzell P, Hägg O, Borgström F. Cost of low back pain: results from a national register study in Sweden. Eur Spine J 2018;27(11):2875−81. https://doi.org/10.1007/s00586-018-5742-6.

[9] McIntosh G, Steenstra I, Hogg-Johnson S, Carter T, Hall H. Lack of prognostic model validation in low back pain prediction studies: a systematic review. Clin J Pain 2018;34(8):748−54.

[10] Lubelski D, Hersh A, Azad TD, Ehresman J, Pennington Z, Lehner K, Sciubba DM. Prediction models in degenerative spine surgery: a systematic review. Global Spine J 2021;11(1_suppl):79s−88s. https://doi.org/10.1177/2192568220959037.

[11] Silva FG, Costa LO, Hancock MJ, Palomo GA, Costa LC, da Silva T. No prognostic model for people with recent-onset low back pain has yet been demonstrated to be suitable for use in clinical practice: a systematic review. J Physiother 2022;68(2):99−109. https://doi.org/10.1016/j.jphys.2022.03.009.

[12] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. https://doi.org/10.1136/bmj.n71.

[13] Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One 2012;7(2):e32844. https://doi.org/10.1371/journal.pone.0032844.

[14] Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis research strategy (PROGRESS) 2: prognostic factor research. PLoS Med 2013;10(2):e1001380. https://doi.org/10.1371/journal.pmed.1001380.

[15] Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11(10):e1001744. https://doi.org/10.1371/journal.pmed.1001744.

[16] Debray TPA, JAAG Damen, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460. https://doi.org/10.1136/bmj.i6460.

[17] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51−8. https://doi.org/10.7326/M18-1376.

[18] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128−38. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

[19] Knoop J, van Lankveld W, Beijer L, Geerdink FJB, Heymans MW, Hoogeboom TJ, et al. Development and internal validation of a machine learning prediction model for low back pain non-recovery in patients with an acute episode consulting a physiotherapist in primary care. BMC Musculoskelet Disord 2022;23(1):834. https://doi.org/10.1186/s12891-022-05718-7.

[20] Kovacs FM, Seco J, Royuela A, Corcoll Reixach J, Abraira V, Spanish Back Pain Research Network. Predicting the evolution of low back pain patients in routine clinical practice: results from a registry within the Spanish National Health Service. Spine J 2012;12(11):1008−20. https://doi.org/10.1016/j.spinee.2012.10.007.

[21] Larivière C, Rabhi K, Preuss R, Coutu MF, Roy N, Henry SM. Derivation of clinical prediction rules for identifying patients with non-acute low back pain who respond best to a lumbar stabilization exercise program at post-treatment and six-month follow-up. PLoS One 2022;17(4):e0265970. https://doi.org/10.1371/journal.pone.0265970.

[22] Stolze LR, Allison SC, Childs JD. Derivation of a preliminary clinical prediction rule for identifying a subgroup of patients with low back pain likely to benefit from Pilates-based exercise. J Orthop Sports Phys Ther 2012;42(5):425−36. https://doi.org/10.2519/jospt.2012.3826.

[23] van Hooff ML, Spruit M, O'Dowd JK, van Lankveld W, Fairbank JC, van Limbeek J. Predictive factors for successful clinical outcome 1 year after an intensive combined physical and psychological programme for chronic low back pain. Eur Spine J Jan 2014;23(1):102−12. https://doi.org/10.1007/s00586-013-2844-z.

[24] Heymans MW, van Buuren S, Knol DL, Anema JR, van Mechelen W, de Vet HC. The prognosis of chronic low back pain is determined by changes in pain and disability in the initial period. Spine J 2010;10(10):847−56. https://doi.org/10.1016/j.spinee.2010.06.005.

[25] Zhao P, Xue J, Xu X, Wang L, Chen D. Logistic model and gradient boosting machine model for physical therapy of lumbar disc herniation. Comput Times 2022;2022:4799248. https://doi.org/10.1155/2022/4799248.

[26] Liu P, Wu Y, Xiao Z, Gold LS, Heagerty PJ, Annaswamy T, et al. Estimating individualized treatment effects using a risk-modeling approach: an application to epidural steroid injections for lumbar spinal stenosis. Pain 2023;164(4).

[27] Childs JD, Fritz JM, Flynn TW, Irrgang JJ, Johnson KK, Majkowski GR, et al. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: a validation study. Ann Intern Med 2004;141(12):920−8.

[28] Traeger AC, Henschke N, Hübscher M, Williams CM, Kamper SJ, Maher CG, et al. Estimating the risk of chronic pain: development and validation of a prognostic model (PICKUP) for patients with acute low back pain. PLoS Med 2016;13(5):e1002019. https://doi.org/10.1371/journal.pmed.1002019.

[29] Gregg CD, McIntosh G, Hall H, Hoffman CW. Prognostic factors associated with low back pain outcomes. J Prim Health Care 2014;6(1):23−30.

[30] Heymans MW, Anema JR, van Buuren S, Knol DL, van Mechelen W, de Vet HCW. Return to work in a cohort of low back pain patients: development and validation of a clinical prediction rule. J Occup Rehabil 2009;19(2):155−65.

[31] Heymans MW, Ford JJ, McMeeken JM, Chan A, de Vet HC, van Mechelen W. Exploring the contribution of patient-reported and clinician based variables for the prediction of low back work status. J Occup Rehabil 2007;17(3):383−97.

[32] Law R, Lee E, Law SW, Chan B, Chen PP, Szeto G. The predictive validity of OMPQ on the rehabilitation outcomes for patients with acute and subacute non-specific LBP in a Chinese population. J Occup Rehabil 2013;23(3):361−70. https://doi.org/10.1007/s10926-012-9404-y.

[33] Leung GCN, Cheung PWH, Lau G, Lau ST, Luk KDK, Wong YW, et al. Multidisciplinary programme for rehabilitation of chronic low back pain - factors predicting successful return to work. BMC

Musculoskelet Disord 2021;22(1):251. https://doi.org/10.1186/s12891-021-04122-x.

[34] Lindell O, Johansson SE, Strender LE. Predictors of stable return-to-work in non-acute, non-specific spinal pain: low total prior sick-listing, high self prediction and young age. A two-year prospective cohort study. BMC Fam Pract 2010;11:53. https://doi.org/10.1186/1471-2296-11-53.

[35] Opsommer E, Rivier G, Crombez G, Hilfiker R. The predictive value of subsets of the Orebro Musculoskeletal Pain Screening Questionnaire for return to work in chronic low back pain. Eur J Phys Rehabil Med 2017;53(3):359−65. https://doi.org/10.23736/S1973-9087.17.04398-2.

[36] Axen I, Jones JJ, Rosenbaum A, Lövgren PW, Halasz L, Larsen K, Leboeuf-Yde C. The Nordic Back Pain Subpopulation Program: validation and improvement of a predictive model for treatment outcome in patients with low back pain receiving chiropractic treatment. J Manip Physiol Ther 2005;28(6):381−5.

[37] Kendell M, Beales D, O'Sullivan P, Rabey M, Hill J, Smith A. The predictive ability of the STarT Back Tool was limited in people with chronic low back pain: a prospective cohort study. J Physiother 2018;64(2):107−13. https://doi.org/10.1016/j.jphys.2018.02.009.

[38] Kongsted A, Vach W, Axo M, Bech RN, Hestbaek L. Expectation of recovery from low back pain: a longitudinal cohort study investigating patient characteristics related to expectations and the association between expectations and 3-month outcome. Spine 2014;39(1):81−90. https://doi.org/10.1097/BRS.0000000000000059.

[39] Malmqvist S, Leboeuf-Yde C, Ahola T, Andersson O, Ekström K, Pekkarinen H, et al. The Nordic back pain subpopulation program: predicting outcome among chiropractic patients in Finland. Chiropr Osteopathy 2008;16:13. https://doi.org/10.1186/1746-1340-16-13.

[40] Szita J, Kiss L, Biczo A, Feher K, Varga PP, Lazary A. Outcome of group physical therapy treatment for non-specific low back pain patients can be predicted with the cross-culturally adapted and validated Hungarian version STarT back screening tool. Disabil Rehabil 2022;44(8):1427−35. https://doi.org/10.1080/09638288.2020.1799248.

[41] Verkerk K, Luijsterburg PA, Heymans MW, Ronchetti I, Miedema HS, Koes BW, et al. Prognostic factors and course for successful clinical outcome quality of life and patients' perceived effect after a cognitive behavior therapy for chronic non-specific low back pain: a 12-months prospective study. Manual Ther 2015;20(1):96−102. https://doi.org/10.1016/j.math.2014.07.003.

[42] Leboeuf-Yde C, Rosenbaum A, Axen I, et al. The Nordic subpopulation research programme: prediction of treatment outcome in patients with low-back pain treated by chiropractors-does the psychological pro. J Am Chiropr Assoc 2010;47(1):20−1.

[43] Lubelski D, Thompson NR, Agrawal B, Abdullah KG, Alvin MD, Khalaf T, et al. Prediction of quality of life improvements in patients with lumbar stenosis following use of membrane stabilizing agents. Clin Neurol Neurosurg 2015;139:234−40. https://doi.org/10.1016/j.clineuro.2015.10.018.

[44] Ramakrishnan A, Michael Webb K, Cowperthwaite MC. One-year outcomes of early-crossover patients in a cohort receiving nonoperative care for lumbar disc herniation. J Neurosurg Spine 2017;27(4):391−6. https://doi.org/10.3171/2017.2.SPINE16760.

[45] Gross DP, Battié MC. Predicting timely recovery and recurrence following multidisciplinary rehabilitation in patients with compensated low back pain. Spine 2005;30(2):235−40. https://doi.org/10.1097/01.brs.0000150485.51681.80.

[46] Fu Y, Feller D, Koes B, Chiarotto A. Prognostic models for chronic low back pain outcomes in primary care are at high risk of bias and lack validation-high quality studies are needed: a systematic review. J Orthop Sports Phys Ther 2024;54:1−36. https://doi.org/10.2519/jospt.2024.12081.

[47] Zhao P, Xue J, Xu X, Wang L, Chen D. Logistic model and gradient boosting machine model for physical therapy of lumbar disc herniation. Comput Math Methods Med 2022;2022:4799248. https://doi.org/10.1155/2022/4799248.

[48] Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348:g1687. https://doi.org/10.1136/bmj.g1687.

[49] Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagn Progn Res 2019;3(1):16. https://doi.org/10.1186/s41512-019-0060-y.