Challenges and solutions in prognostic prediction models in spinal disorders

Wingbermuhle, Roel; Chiarotto, Alessandro; Koes, Bart; Heymans, Martijn W; van Trijffel, Emiel

published in
Journal of Clinical Epidemiology
2021

DOI (link to publisher)
10.1016/j.jclinepi.2020.12.017

document version
Publisher's PDF, also known as Version of record

Journal of Clinical Epidemiology 132 (2021) 125−130

Journal of Clinical Epidemiology

COMMENTARY

Challenges and solutions in prognostic prediction models in spinal disorders

Roel W. Wingbermühle[a,b,*], Alessandro Chiarotto[b,c], Bart Koes[b,d], Martijn W. Heymans[e], Emiel van Trijffel[a,f]

[a]*SOMT University of Physiotherapy, Amersfoort, The Netherlands*
[b]*Department of General Practice, Erasmus MC, University Medical Center, Rotterdam, The Netherlands*
[c]*Department of Health Sciences, Faculty of Science, VU University, Amsterdam Movement Sciences, Amsterdam, The Netherlands*
[d]*Center for Muscle and Joint Health, University of Southern Denmark, Odense M, Denmark*
[e]*Department of Epidemiology and Data Science, Amsterdam University Medical Center, Amsterdam, The Netherlands*
[f]*Experimental Anatomy Research Department, Department of Physiotherapy, Human physiology and Anatomy, Faculty of Physical Education and Physiotherapy, Vrije Universiteit Brussels, Brussels, Belgium*

Accepted 14 December 2020; Published online 18 January 2021

Abstract

Methodological shortcomings in prognostic modeling for patients with spinal disorders are highly common. This general commentary discusses methodological challenges related to the specific nature of this field. Five specific methodological challenges in prognostic modeling for patients with spinal disorders are presented with their potential solutions, as related to the choice of study participants, purpose of studies, limitations in measurements of outcomes and predictors, complexity of recovery predictions, and confusion of prognosis and treatment response. Large studies specifically designed for prognostic model research are needed, using standard baseline measurement sets, clearly describing participants' recruitment and accounting and correcting for measurement limitations.

Keywords: Prognostic models; Spinal disorders; Methodological challenges; Methodological shortcomings; Clinical prediction models; Musculoskeletal disorders

1. Introduction

Prediction models estimate the probability of a condition being present or a future health outcome to occur by combining values of multiple predictors. In clinical practice, prediction models aim to improve the quality of care for *individual* patients by supporting decisions on prevention, diagnosis (diagnostic models), prognosis (prognostic models), or treatment (predictive models) [1]. In this commentary, we focus on methodological challenges and possible methodological improvements of prognostic prediction models for spinal disorders, based on existing evidence about prognostic modeling and on our own research experience in the field. Studies of prognostic models comprise three consecutive stages: model development (derivation), preferably with internal validation; validation in new settings (external validation); and assessment of a model's clinical impact [2]. The shift to personalized medicine has led to a vast amount of published prognostic models, including an increasing number of studies in the spinal field [3,4].

Worldwide, low back pain (LBP) and neck pain (NP) are major health problems and leading causes of disability [5]. These spinal disorders may concern specific diseases (e.g., spinal stenosis, axial spondyloarthritis, malignancy, fracture); however, the vast majority concern conditions without an identifiable pathoanatomical cause are thus labeled as nonspecific. LBP and NP are increasingly understood as complex conditions with a variable course of related episodes and multiple interacting biopsychosocial contributors [6,7]. After an initial improvement in pain and functioning, their long-

Declaration of interests: None.

The submitted material has not been published and is not under consideration for publication elsewhere.

* Corresponding author. SOMT University of Physiotherapy, P.O. Box 585, 3800 AN Amersfoort, The Netherlands. Tel.: +31 334560737.

E-mail address: r.wingbermuhle@somtuniversity.nl (R.W. Wingbermühle).

https://doi.org/10.1016/j.jclinepi.2020.12.017
0895-4356/

What is new?

Key findings

- An increasing number of prognostic model studies are published in the field of spinal disorders. However, methodological shortcomings in these studies are highly common. Five methodological challenges related to the specific nature of the field are described, and potential solutions are suggested.

What this adds to what was known?

- Specific methodological challenges in prognostic modeling for patients with spinal disorders as related to the choice of study participants, purpose of studies, heterogeneity of outcome measurements, limitations in measurements of outcomes and predictors, complexity of recovery predictions, and confusion of prognosis and treatment response are presented, illustrated, and discussed, and potential solutions are suggested.

What is the implication and what should change now?

- New, large studies are needed specifically designed for prognostic model research, using standard baseline measurement sets, clearly describing participants' recruitment with a sharp demarcation of subgroups with expected different prognosis and accounting and correcting for measurement limitations.

term clinical course is unfavorable in a substantial proportion of people [8,9]. Prognostic models intend to distinguish patients with an unfavorable long-term course from those with a favorable course, have the potential to decrease patients' burden, and can make contributions to cost-effective health care. Early comprehensive treatment given to people with a favorable short-term course is unnecessary and probably not cost-effective. Interventions in this group with a favorable prognosis can even be contraeffective. However, early identification of people at high risk of an unfavorable long-term outcome can be beneficial, as this enables clinicians to provide appropriate advice and cost-effective treatments [8−11]. In this commentary, specific methodological shortcomings in the research of prognostic modeling for patients with spinal disorders are presented and discussed, and potential solutions are suggested.

2. Spinal prognostic model studies

2.1. Methodological shortcomings in general

Common methodological shortcomings in prognostic modeling such as inadequate sample size compared with the number of candidate predictor categories, predictor selection based purely on statistical significance, categorization of continuous predictors, and lack of reporting of key performance measures and poor overall reporting have also been identified in the spinal field [3,12]. These pitfalls often lead to models that are overfitted and overoptimistic or to model predictors that reflect chance or biased associations with the outcome, resulting in models that generalize poorly to other clinical settings and patients [13]. Moreover, prognostic models for spinal disorders do not typically reach their validation phase, and impact studies are absent [3,4]. These common shortcomings can to a large extent be addressed by following currently available methodological standards for designing, executing, and reporting prediction models in health care [12,14].

2.2. Specific methodological challenges

2.2.1. Challenge 1: problems with choice of participants

The adoption of different inclusion or exclusion criteria across models may result in different models that are difficult to compare. In addition, the adoption of unclear criteria may lead to models not applicable to the initial target population, which limits generalizability. For example, in a systematic review on prognostic models for NP, the original studies included participants based on highly variable and sometimes unclear NP criteria [3]. One study included people with whiplash-associated disorders (WADs) Grade I and II, whereas another concerned people with WAD Grade II and III [3]. WAD I reflects NP and perceived stiffness, WAD II includes the presence of physical signs, and WAD III includes neurological signs. It is known that patients with WAD III have a different prognosis, which makes it hard to compare these prognostic models [15].

To counter this problem, there should be a clear description of recruitment and selection criteria, with demarcation of subgroups with expected different prognosis (e.g., WAD Grade III) is recommended.

2.2.2. Challenge 2: use of studies not purposively designed for prognostic models

As spinal pain is mostly diagnosed as nonspecific, the focus in this field is on functional health relating to common signs and symptoms that are mainly identified through history taking, physical examination, and patient-reported questionnaires. These predictors mostly result in models with limited predictive performance [3]. For example, we developed models for NP recovery in an existing patient cohort with potential model predictors selected from the literature and clinical perspective. The immediate posttreatment recovery models showed optimism adjusted Nagelkerke R^2 of 0.09 (interquartile range [IQR] 0.08−0.11), 0.09 (IQR 0.07−0.11), and 0.21 (IQR 0.19−0.23) for pain, perceived improvement, and disability, respectively. The models for 1-year recovery displayed an R^2 of 0.06 (IQR 0.05−0.07), 0.07 (IQR 0.06−0.08), and 0.06 (IQR

0.05−0.07) for the same outcomes (submitted). The reason for this limited performance was that the cohort used for the development of this model was not originally designed to develop a prognostic model, and many baseline variables were not operationalized in an adequate manner to be entered as predictors, leading to a poor model's global performance.

The field is also strongly focused on examining patient-reported psychosocial factors, whereas the use of objective markers (e.g., imaging) is consistently discouraged by international guidelines, as these have not proven to add useful diagnostic or prognostic information. The result is that more objective markers are only rarely investigated in cohort studies or clinical trials used to develop prognostic models in the spinal field.

Large data sets purposively designed for prognostic model development or validation can contain a large array of candidate predictors. To develop prognostic models for spinal disorders, researchers should use cohort studies in which a broad range of biological, physical, and psychosocial measures are included. To support this, a recent exploratory prognostic factor study found that three "biological" features seldom evaluated, that is, morning stiffness, painful spinal rotation, and multilevel radiographic osteophytes, predicted long-term LBP in older adults [16]. In addition, to overcome limited information on potential key predictors, the development of baseline standard sets of subjective and objective potential predictors may facilitate measurement and assessment of the most relevant ones. Since 2014, a minimal baseline set for chronic LBP exists, which includes demographic items, medical history, and self-report of symptoms and function [17]. However, there is no evidence on the use of this minimal baseline set so far. An international and multidisciplinary consortium may focus on developing a standard set of potential prognostic factors to be measured in cohort studies in the field of spinal disorders. This would facilitate the development of cross-cohort prognostic models and the cross-cultural external validation of models.

2.2.3. Challenge 3: limitations in measurement of outcomes and predictors

Health constructs such as pain intensity, physical functioning, perceived treatment effect, or health-related quality of life represent core outcomes in patients with spinal disorders [18]. These are also the most frequently used recovery outcomes in prognostic research. Nevertheless, the definition of recovery can vary substantially across studies. Some studies may define recovery in terms of pain reduction, whereas other studies may define it as an improvement in physical functioning. These discrepancies highlight the uncertainty around the recovery concept, which is often multidimensional from a patient perspective [19]. The aforementioned core outcomes are mainly measured with Patient-Reported Outcome Measures (PROMs). Nevertheless, different PROMs can measure the same construct, and if these PROMs are not truly measuring the same construct, they may result in models including different predictors. Our systematic review on prognostic models for NP confirmed that a large variety of PROMs and cutoffs are used [3]. For example, for (neck-related) physical functioning, the Northwick Park Questionnaire, the Pain Disability Index, the Neck Pain Outcome Score, a Visual Analogue Scale for daily activities, or the Neck Disability Index (NDI) were used. In addition, cutoffs for NDI varied from 5/50 to 15/100 or 8% to 10% [3]. To deal with this heterogeneity, the development of consensus-based core outcome sets for prognostic models in spinal disorders may be a solution.

It should also be noted that, although PROMs can be used as continuous measures, their scores are often dichotomized in prognostic modeling to compare recovered (or improved) versus nonrecovered (or nonimproved) patients. Parameters indicating a minimal improvement that patients would consider as important, such as the Minimal Important Change (MIC), can be used to dichotomize PROMs and various methods exist to calculate these parameters. An alternative method to determine recovery is to use a percentage of improvement (e.g., 30%, 50%) based on consensus among experts [20]. Nevertheless, the use of different threshold parameters to define outcomes can lead to the selection of different predictors in a model [21]. A solution to this issue may be to adopt recent methodological development on the MIC estimation. For example, a predictive approach to calculate the MIC was found to be more precise than the standard anchor-based approach, as it can more easily adjust for baseline scores and for the number of improved patients [22].

Physical tests, performance tests, biomarkers, and other measures can be used to measure predictors. PROMs are probably also the most frequently used instruments for measuring potential predictors, but they are not free from bias and error [23]. Here we briefly discuss some measurement limitations that afflict PROMs on three key measurement properties: content validity, structural validity, and measurement error.

Content validity concerns whether a measure is an adequate reflection of the construct to be measured. However, it is only rarely evaluated in spinal disorders. A systematic review on the content validity of 17 PROMs used to measure physical functioning in LBP found high-quality evidence for only one PROM [24]. Including PROMs with unknown content validity may lead to prognostic models that do not adequately reflect the constructs that are meant to be measured. Therefore, PROMs with high-quality evidence for satisfactory content validity should be preferred.

Structural validity—which refers to the degree to which the dimensionality of a measure is an adequate reflection of the dimensionality of the measured construct—is often problematic for widely used measures, which may not be unidimensional for constructs, such as pain, disability, or

health-related quality of life. For instance, widely used PROMs in patients with LBP have displayed poor or conflicting unidimensionality [24]. Introducing patient-rated predictor and outcome measures with poor or uncertain dimensionality in prognostic modeling may introduce biased models. One solution to mitigate this issue is to use Item Response Theory (IRT)-based scores instead of the standard used sum-based scores. A large variety of IRT models is available to model the ''real'' dimensionality of a PROM and to provide scores that take that dimensionality into account. A comparison of IRT-based versus sum-based scores showed that IRT-based scores provide more precise estimates of longitudinal data analyses of PROMs [25].

Measurement error and misclassification of predictors and outcome is poorly addressed in medical research [26]. One parameter often used to assess measurement error of PROMs is the Smallest Detectable Change, which refers to a patient's score beyond which ''true'' changes in the construct to be measured are reflected. In patients with LBP, for instance, Smallest Detectable Change of the Roland-Morris Disability Questionnaire and Oswestry Disability Index vary substantially from 4.0 to 8.6 points (0−24 scale) and 11.0 to 16.7 points (0−100 scale), respectively [27]. Measurement error of self-reported predictors, such as height and weight, appears to influence model performance; random error decreases calibration and discrimination, whereas systematic error affects calibration and does not influence discrimination [28]. Studies are needed (e.g., simulation studies) to investigate the influence and impact of measurement error and misclassification—for predictors and outcomes—of commonly used PROMs on spinal model performance. Subsequently, researchers may correct for these errors, if possible, using ancillary studies and adjustment analysis methods (e.g., regression calibration, simulation-extrapolation, latent variable models), performing sensitivity analyses, or deciding to use alternative measures [26].

2.2.4. Challenge 4: predicting recovery from spinal disorders is complex

Nonspecific spinal disorders can typically be regarded as complex health problems with many interacting factors contributing to pain and disability [6,29]. Consequently, predicting long-term outcomes such as recovery undoubtedly has a complex nature. Consequently, current approaches to building models may not adequately cover the many, often perhaps unknown variables and their interactions involved that also may change dynamically over time.

Only very few studies consider predictors trajectory over time and interactions during their model building. For instance, Schellingerhout et al. [30] found that ''accompanying headache'' interacted with four clinical features to predict persistent neck complaints. Bohman et al. [31] included a factor-by-time interaction term in their NDI model that showed an area under the curve (95% confidence interval) of 0.67 (0.59−0.75) after internal validation. However, they based the time factor on all follow-ups, which limits model's clinical utility. Heymans et al. included clinically relevant change in pain intensity and disability status in their model predicting chronic LBP, which showed good performance with an area under the curve of 0.80 and explained variation of 37% after internal validation [32]. Changed predictor scores were calculated from baseline over 3 months, which limits the model's clinical utility.

We hypothesize that including interaction and predictor trajectory over time variables, defined a priori based on plausible biological mechanisms, during model building has the potential to improve the prediction of outcomes in patients with spinal disorders.

2.2.5. Challenge 5: confusion in prognostic factors and predictors of treatment response

Prognostic factors do not necessarily also predict the effect of treatment. It is also important to note that models that predict the treatment effect require different study designs (i.e., randomized clinical trials) compared with models that predict outcome in general, regardless of treatment applied (i.e., cohort studies) [1]. Many studies in the spinal field use designs that cannot validly differentiate between predictors of treatment effect and general outcome [1]. Predictors of treatment effect are evaluated by investigating the interaction of that predictor with treatment as additional effect on the outcome [1]. Single-arm cohort studies may provide exploratory information and hypotheses on candidate factors for influencing treatment effect, but further double-arm trials are needed for stronger model development and validation [1].

3. Conclusion

Clear description of participants' recruitment and selection is paramount in spinal research prognostic models, with a sharp demarcation of subgroups with expected different prognosis and a clear message about models intended use. There is a need for studies to investigate the influence and impact of measurement limitations that afflict widely used PROMs on key properties as content validity, structural validity, and measurement error, allowing researchers to account and correct for these measurement limitations. Several problems in spinal prognostic modeling can be alleviated if large studies specifically designed for prognostic model research are designed preferably using baseline standard measurement sets that are tuned to cover a wide array of biological, physical, and psychosocial measures. New methods for analyzing complex networks of interacting variables may be promising solutions to account for the complex nature of spinal disorders. We envision that machine learning techniques will be capable of discovering

and modeling prognostic factors and their interactions, dynamically and in real time, in large linked data sets from local electronic health care records, wearables, and social and genomic information of patients with nonspecific spinal pain. Although artificial intelligence has so far sparsely been used in nonspecific LBP prognosis in only small data sets [33], several machine learning algorithms have been developed for clinical prognostication after spinal surgery [34]. In addition, in line with recent guidance on prognostic modeling methodology, performance measures such as calibration and discrimination should be reported both in derivation and validation studies, where net benefit and decision curves can additionally capture model's clinical usefulness [12].

References

[1] Riley RD, Van Der Windt DA, Croft P, Moons KGM. Prognosis research in healthcare. First. Oxford, England: Oxford University Press; 2019.

[2] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2012;10(2):e1001381.

[3] Wingbermühle RW, van Trijffel E, Nelissen PM, Koes B, Verhagen AP. Few promising multivariable prognostic models exist for recovery of people with non-specific neck pain in musculoskeletal primary care: a systematic review. J Physiother 2018;64(1):16−23.

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

[5] James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990−2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392:1789−858.

[6] Hartvigsen J, Hancock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, et al. What low back pain is and why we need to pay attention. Lancet 20186736.

[7] Guzman J, Hurwitz EL, Carroll LJ, Haldeman S, Côté P, Carragee EJ, et al. A new conceptual model of neck pain: linking onset, course, and care: the bone and joint decade 2000-2010 task force on neck pain and its associated disorders. Spine (Phila Pa 1976) 2008;33(4 Suppl):S14−23.

[8] Traeger AC, Hübscher M, McAuley JH. Understanding the usefulness of prognostic models in clinical decision-making. J Physiother 2017;63(2):121−5.

[9] Hush JM, Lin CC, Michaleff ZA, Verhagen A, Refshauge KM. Prognosis of acute idiopathic neck pain is poor: a systematic review and meta-analysis. Arch Phys Med Rehabil 2011;92(5):824−9.

[10] Buchbinder R, van Tulder M, Öberg B, Costa LM, Woolf A, Schoene M, et al. Low back pain: a call for action. Lancet 2018; 391(10137):2384−8.

[11] da C Menezes Costa L, Maher CG, Hancock MJ, McAuley JH, Herbert RD, Costa LO. The prognosis of acute and persistent low-back pain: a meta-analysis. CMAJ 2012;184(11):E613−24.

[12] Haskins R, Rivett DA, Osmotherly PG. Clinical prediction rules in the physiotherapy management of low back pain: a systematic review. Man Ther 2012;17(1):9−21.

[13] Steyerberg EW, Uno H, Ioannidis JPA, van Calster B, Ukaegbu C, Dhingra T, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol 2018;98: 133−43.

[14] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162(1):W1−73.

[15] Côté P, Cassidy JD, Carroll L, Frank JW, Bombardier C. A systematic review of the prognosis of acute whiplash and a new conceptual framework to synthesize the literature. Spine (Phila Pa 1976) 2001; 26(19):E445−58.

[16] van den Berg R, Chiarotto A, Enthoven WT, de Schepper E, Oei EHG, Koes BW, et al. Clinical and radiographic features of spinal osteoarthritis predict long-term persistence and severity of back pain in older adults. Ann Phys Rehabil Med 20201735−45.

[17] Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, et al. Report of the NIH task force on research standards for chronic low back pain. J Pain 2014;15:569−85.

[18] Chiarotto A, Deyo RA, Terwee CB, Boers M, Buchbinder R, Corbin TP, et al. Core outcome domains for clinical trials in nonspecific low back pain. Eur Spine J 2015;24(6):1127−42.

[19] Hush JM, Refshauge K, Sullivan G, De Souza L, Maher CG, McAuley JH. Recovery: what does this mean to patients with low back pain? Arthritis Rheum 2008;61(1):124−31.

[20] Ostelo R, Stratford Deyo P, Waddell G, Croft P, Von Korff M, Bouter LM, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine (Phila Pa 1976) 2008;33(1):90−4.

[21] Schwind J, Learman K, O'Halloran B, Showalter C, Cook C. Different minimally important clinical difference (MCID) scores lead to different clinical prediction rules for the Oswestry disability index for the same sample of patients. J Man Manip Ther 2013; 21(2):71−8.

[22] Terluin B, Eekhout I, Terwee CB, de Vet HC. Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. J Clin Epidemiol 2015; 68(12):1388−96.

[23] Chiarotto A. Patient-reported outcome measures: best is the enemy of good (but what if good is not good enough?). J Orthop Sports Phys Ther 2019;49(2):39−42.

[24] Chiarotto A, Ostelo RW, Boers M, Terwee CB. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. J Clin Epidemiol 2018;95:73−93.

[25] Gorter R, Fox JP, Twisk JW. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Med Res Methodol 2015;15:1−12.

[26] Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. J Clin Epidemiol 2018;98:89−97.

[27] Chiarotto A, Maxwell LJ, Terwee CB, Wells GA, Tugwell P, Ostelo RW. Roland-Morris Disability Questionnaire and Oswestry Disability Index: which has better measurement properties for measuring physical functioning in nonspecific low back pain? Systematic review and meta-analysis. Phys Ther 2016;96:1620−37.

[28] Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. Popul Health Metr 2012;10(1):20.

[29] Sturmberg JP. The value of systems and complexity sciences of healthcare. Cham, Switzerland: Springer International Publishing; 2016.

[30] Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HC, Koes BW. Prognosis of patients with nonspecific neck pain: development and external validation of a prediction rule for persistence of complaints. Spine (Phila Pa 1976) 2010;35(17):E827−35.

[31] Bohman T, Bottai M, Björklund M. Predictive models for short-term and long-term improvement in women under physiotherapy for chronic disabling neck pain: a longitudinal cohort study. BMJ Open 2019;9(4):e024557.

[32] 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.

[33] Tagliaferri SD, Angelova M, Zhao X, Owen PJ, Miller CT, Wilkin T, et al. Artificial intelligence to improve back pain outcomes and lessons learnt from clinical classification approaches: three systematic reviews. NPJ Digit Med 2020; 3(1):93.

[34] Schwartz JT, Gao M, Geng EA, Mody KS, Mikhail CM, Cho SK. Applications of machine learning using electronic medical records in spine surgery. Neurospine 2019;16(4):643−53.