Physiotherapy 113 (2021) 61–72

Development and internal validation of prognostic models for recovery in patients with non-specific neck pain presenting in primary care

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

[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, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, The Netherlands*
[d] *Vrije Universiteit Brussels, Experimental Anatomy Research Department, Department of Physiotherapy, Human Physiology and Anatomy, Faculty of Physical Education and Physiotherapy, Brussels, Belgium*
[e] *Center for Muscle and Joint Health, University of Southern Denmark, Odense M, Denmark*
[f] *University of Technology Sydney, Level 4 Building 7, Sydney, Australia*
[g] *Department of Epidemiology and Data Science, VU University Medical Center, Amsterdam, The Netherlands*

Abstract

Objectives Development and internal validation of prognostic models for post-treatment and 1-year recovery in patients with neck pain in primary care.
Design Prospective cohort study.
Setting Primary care manual therapy practices.
Participants Patients with non-specific neck pain of any duration ($n = 1193$).
Intervention Usual care manual therapy.
Outcome measures Recovery defined in terms of pain intensity, disability, and global perceived improvement directly post-treatment and at 1-year follow-up.
Results All post-treatment models exhibited acceptable discriminative performance after derivation (AUC $\geq$ 0.7). The developed post-treatment disability model exhibited the best overall performance ($R^2 = 0.24$; IQR, 0.22–0.26), discrimination (AUC = 0.75; 95% CI, 0.63–0.84), and calibration (slope 0.92; IQR, 0.91–0.93). After internal validation and penalization, this model retained acceptable discriminative performance (AUC = 0.74). The five other models, including those predicting 1-year recovery, did not reach acceptable discriminative performance after internal validation. Baseline pain duration, disability, and pain intensity were consistent predictors across models.
Conclusion A post-treatment prognostic model for disability was successfully developed and internally validated. This model has potential to inform primary care clinicians about a patient's individual prognosis after treatment, but external validation is required before clinical use can be recommended.

Contribution of the paper

- Existing prognostic models for patients with non-specific neck pain present substantial methodological shortcomings, which prevent their clinical use.
- We developed and internally validated prognostic models to predict recovery in patients with neck pain.
- The prognostic model for post-treatment disability exhibited good performance and calibration, showing promise for external validation and clinical use.

Keywords: Neck pain; Recovery; Prognostic models; Model development; Prognostic factors

[*] Corresponding author at: SOMT University of Physiotherapy, P.O. Box 585, 3800 AN Amersfoort, The Netherlands.
E-mail addresses: r.wingbermuhle@somtuniversity.nl, roelwingbermuhle@me.com (R.W. Wingbermühle).

https://doi.org/10.1016/j.physio.2021.05.011

Introduction

Neck pain is a top five cause of Years Lived with Disability in high and middle income countries and, after low back pain, the second worldwide largest cause of musculoskeletal disability [1]. Recovery from non-specific neck pain mainly takes place in the first six weeks with very little further long-term improvement of pain and disability [2,3]. The prevalence of chronic neck pain, i.e. pain lasting longer than three months, has increased from 2005 to 2015 by 21% up to approximately 358 million people worldwide and it is likely to increase further in Western countries due to an ageing population [4]. Noninvasive primary care interventions (e.g. mobilisations and manipulations, exercise, psychosocial interventions, or combinations) are reported as effective treatments for non-specific neck pain [5–7].

An accurate individual prognosis at intake can inform clinicians and patients in shared clinical decisions [8]. For example, in patients with a high risk of poor prognosis, subsequent effective treatment interventions may improve the patients' prognosis; at the same time, a wait-and-see approach in patients with a very low risk of poor prognosis can limit exposure to unnecessary treatments and reduce costs [8]. Separate prognostic factors which are consistently reported for outcomes on neck-related pain, physical functioning, and perceived recovery are: age, sex, baseline pain intensity, baseline disability, and past history of neck pain [9–11]. Prognostic prediction models (in short: prognostic models) provide probabilities for patients based on their individual combination of predictor values and can support clinicians in their clinical decisions [12]. Prognostic models have been shown to improve prognostic accuracy in various healthcare fields [13,14]. However, a recent systematic review concluded that the clinical utility of currently available prognostic models in people with neck pain is limited [15]. Overall, the methodological quality of the studies included in this review was low with the large majority of studies lacking sufficient sample size and internal validation [15]. Furthermore, from the three promising models as defined in the systematic review, two appeared invalid in a subsequent external validation study and a third model specifically focusing on patients with whiplash associated disorders could not be tested [16]. Therefore, there is a need to develop a prognostic model for recovery in patients with neck pain that exhibits satisfactory prediction. This model should be developed in a cohort of patients with adequate sample size, and it should be internally validated.

The aim of this study was to develop and internally validate prognostic models that predict at intake post-treatment and 1-year follow-up recovery of neck pain, disability, and global perceived improvement in patients treated with manual therapy in primary care.

Methods

Design

For this model derivation study, the authors used data from a prospective cohort study, the 'Amersfoorts Nekonderzoek of the Master manuele therapie Opleiding' (ANIMO), conducted from 2007 to 2009. In total, 345 manual therapists in the Netherlands recruited 1311 consecutive patients between 18 and 80 years presenting with non-specific neck pain of any duration. Participants providing baseline data and having signed informed consent were deemed eligible ($n = 1193$). Neck pain with or without associated arm pain was classified as non-specific if the pain could not be attributed to a specific underlying pathology (i.e., no red flags were present). Study characteristics (e.g., setting, inclusion criteria, measurement procedures) have been described in detail elsewhere [17]. Participating patients received usual care multimodal manual therapy which may have included specific joint mobilizations, high velocity thrust techniques, myofascial techniques, giving advice, or specific exercises. Mean treatment duration was 37.9 days, mean number of treatment sessions was 4.3. The Erasmus Medical Centre Ethics Committee Rotterdam, the Netherlands (MEC-2007-359) approved this study.

This study was conducted following the PROGRESS group recommendations [18] and reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement [19].

Candidate model predictors

The authors based selection of candidate predictors for the models on the literature and clinical credibility of variables in combination with their reliability, applicability, and costs [20–23], while avoiding univariable pre-selection [8]. The following predictors were considered: age, sex, previous neck pain episode, neck pain duration (acute 0–6 weeks, sub-acute 6–12 weeks and chronic >12 weeks), pain intensity (measured with a Numerical Rating Scale (*NRS*)), and disability (Neck Disability Index – Dutch version (*NDI-DV*)) [11,24,25]. Furthermore, the authors included six additional candidate predictors regarded in the literature as clinically credible and relatively easy to collect at intake [9,11,25]: accompanying headache (*yes/no),* accompanying low back pain (*yes/no),* accompanying radiating arm pain (*yes/no),* smoking status (*yes/no*), fear-avoidance beliefs *(Fear-Avoidance Beliefs Questionnaire – Dutch version (FABQ-DV)* physical activity subscale [26,27]), and psychological functioning (*Neck Bournemouth Questionnaire-DV (NBQ-DV)* anxiety and depression subscale [28–30]). Additionally, the authors considered other potentially relevant predictors from a clinical perspective: general sleeping problems (*yes/no*), partaking in sporting activities (*yes/no*), and patients' expectation to change due to treatment (*5-point Likert scale, ranging from 'much better' to 'much worse')* [31].

Outcomes

In this study, recovery was used as an umbrella term for three different constructs and outcome measures, which were: (1) for pain as an NRS *(10-point Likert scale)* score dichotomized into >2 for non-recovery and ≤2 for recovery as the latter is considered as a satisfactory state by patients [32]; (2) for disability, by dichotomizing the NDI-DV (*0–50 scale range*), after values were multiplied by two to yield percentages, into <8% for recovery and ≥8% for non-recovery, which is a threshold used before [33,34]; and (3) for global perceived improvement as Global Perceived Effect (GPE) measured on a *7*-point Likert scale where recovery was defined by response options "completely recovered" or "much improved", while non-recovery by responses "slightly improved", "no change", "slightly worse"," much worse", and "worse than ever" represented non-recovery [35,36] Post-treatment follow-up was measured in ANIMO immediately after a course of treatment and defined as no more than three months after intake, and long-term follow-up was measured after one year from inclusion. Outcome questionnaires were returned by post through provided prepaid envelopes.

Missing values

Missing values were evaluated by comparing patients with and without missing values on relevant predictors and by performing t-tests [37–40]. Missing At Random (MAR) was most plausible based on the data not being MCAR according to compared patients and the performed t-tests. Multiple imputation on predictors as well as outcomes using all predictor and outcome variables was performed [38–41]. The method of Multivariate Imputation by Chained Equations (MICE) procedure with generation of 50 imputed data sets was applied [41]. Regression coefficient estimates and standard errors were pooled according to Rubin's Rules, and model performance measures estimated in each of the 50 completed datasets and then combined [39,42,43].

Statistical analysis

Regression model assumptions such as linear relationship between predictor variables and the outcome were evaluated using restricted cubic splines and multicollinearity (Tolerance > 0.2, Variance Inflation Factor < 3). Variables were coded before entering the regression models and categorical variables were transformed into dummy variables [44–46].

Multivariable logistic regressions were estimated for all the models in the imputed ANIMO datasets as primary analysis. A backward elimination approach with the *P*-value set at <0.157 was used as this corresponds to the Akaike information criterion [43,47]. Overall performance was expressed as Nagelkerke's R^2; calibration was estimated by the calibration slope, calibration curve, and the Hosmer–Lemeshow test; and the Area Under Curve (AUC) of the receiver-operating characteristic Curve (ROC) was calculated for quantifying discriminative performance [8,23]. Perfect discriminative performance has a value of 1 and the authors considered discriminative performance acceptable if AUC was ≥0.7 [48]. The calibration plot is obtained across multiply imputed data sets by the following approach that is commonly used to make a calibration plot. In each imputed dataset the predicted probabilities are determined and used to make 10 groups by using 10 deciles. Within these groups the observed outcomes were divided by the sample size of each group to obtain the predicted probabilities. The agreement between these 10 groups is plotted on the calibration curve and a natural cubic spline curve is plotted between the black dots. The groups and calibration curves of each imputed data set are plotted in the same figure, distinguished by the multiple blue lines and multiple black dots for the groups. This makes it possible to evaluate agreement across multiply imputed data sets. Internal validation of all models was performed with bootstrapping in 250 samples, and repeating all development steps. [49]. The authors corrected the models' regression coefficients with the optimism-adjusted calibration slope value and updated the intercept using an "offset" procedure by calculating the linear predictor with the new regression coefficients fixed [50]. All analyses were performed in IBM SPSS 24.0 and R version 3.4.3.

Sensitivity analyses

In addition, the authors estimated all models and their performance measures on the complete case data as sensitivity analyses to allow comparison of models and performance measures obtained on the imputed data.

Sample size and candidate model predictors

The authors performed *a priori* sample size calculations for each model to decide on the amount of candidate predictor parameters, using the procedure described by Riley *et al.* with a shrinkage of 0.8 and R^2 of 0.1 [51]. The proportion post-treatment non-recovery was 21%, 58%, and 21% for pain intensity, disability, and global perceived improvement, respectively, and after 1 year it was 45%, 62%, and 39%, respectively. This resulted in a maximum amount of candidate predictor categories, depending on these outcome proportions, ranging from 14 to 18. Calculations were made with the pmsamplesize package in R.

Results

Baseline characteristics and candidate model predictors

Patients' baseline characteristics and candidate factors were comparable for complete cases (Supplement 2) and

Table 1
Baseline characteristics and candidate model predictors of patients with non-specific neck pain ($n = 1193$).

Baseline characteristics		Missing *n* (%)
Age (years), mean (SD)	44.7 (13.7)	23 (2)
Gender		7 (1)
Female sex, *n* (%)	823 (69)	
Previous neck pain episode		64 (5)
Yes, *n* (%)	755 (67)	
Neck pain duration		122 (10)
Acute 0 to 6 weeks, *n* (%)	420 (39)	
Subacute 6 to 12 weeks, *n* (%)	138 (13)	
Chronic >12 weeks, *n* (%)	513 (48)	
Pain intensity (*NRS, scale 1 to 10*)[c], mean (SD)	4.8 (2.1)	10 (1)
Disability (*NDI, scale 0 to 50*)[d], median [IQR]	12.0 [8.0 to 17.0]	97 (8)
Accompanying headache		0 (0)
Yes, *n* (%)	707 (59)	
Accompanying low back pain		0 (0)
Yes, *n* (%)	538 (45)	
Accompanying radiating arm pain		0 (0)
Yes, *n* (%)	536 (45)	
Accompanying general sleeping problems		0 (0)
Yes, *n* (%)	337 (28)	
Smoking status		3 (0)
Yes, *n* (%)	300 (25)	
Fear-avoidance believes (*FABQ-PA, scale 0 to 24*)[a], median [IQR]	11.0 [6.0 to 15.0]	85 (7)
Emotional functioning (*NBQ-AD, scale 0 to 20*)[b], median [IQR]	7.0 [3.0 to 10.0]	16 (1)
Partaking in sporting activities		4 (0)
Yes, *n* (%)	783 (66)	
Patients' expectation to change due to treatment		3 (0)
Much better, *n* (%)	517 (43)	
Better, *n* (%)	662 (56)	
No change, *n* (%)	10 (1)	
Worse, *n* (%)	1 (0)	
Much worse, *n* (%)	0 (0)	

% rounded up to closest integer.

[a] FABQ-PA = fear-avoidance beliefs questionnaire, physical activity subscale (scale 0–24).

[b] NBQ-AD = Neck Bournemouth Questionnaire, anxiety and depression subscale (scale 0–20), sum score of 11-point numeric subscale of items 4 and 5.

[c] NRS = numeric rating scale.

[d] NDI = neck disability index.

cases with no outcome data (Table 1). Mean age of patients was 44.7 (SD 13.7) years, 69% ($n = 823$) were female, and 67% ($n = 755$) experienced a previous episode and 48% ($n = 513$) was classified as chronic. Mean baseline pain intensity was 4.8 (SD 2.1) and median disability was 12.0 [IQR 8.0–17.0]. The candidate factor for treatment expectations was excluded since it showed an extreme standardised error and coefficient during model estimation.

Outcome values

Outcome values are presented in Table 2. Pain intensity was 2.0 [IQR 1.0–2.0] and 2.8 [IQR 1.0–4.0] post-treatment and at 1-year, respectively. Disability was 5.0 [IQR 1.0–9.0] and 5.0 [IQR 2.0–8.0] post-treatment and at 1-year, respectively.

Missing values

Several baseline characteristics had more than 5% missing values and a few up to 13% (Table 1). The 1-year outcome values reached about 45% missing values and the post-treatment about 55% (Table 2). Baseline characteristics were comparable between complete cases (Supplement 2) and those without outcome data, and the means of several variables differed significantly depending on the missingness of indicator variables, indicating that the MAR assumption is more plausible. Therefore, the authors assumed data were MAR. The authors chose 50 imputed datasets as the rule of thumb is the number of imputations is as large as the percentage of missing data [41]. In fact, the authors had missing data of 46, 43, 43, 53, 54 and 56% in the outcomes. This is on average 42% for all outcomes. The authors applied one run of 50 imputed datasets and developed the different models in the same imputed data to eliminate the influence of missing data imputation on the development of the models. Multicollinearity in the MI model was not checked, but checked between variables before the models were developed. Further, the authors evaluated the convergence plots of the imputed variables and these showed healthy convergence, i.e., no irregular patterns were visible, which is often an indication of that there is no multicollinearity between variables.

Table 2
Pain intensity, disability, and perceived recovery post-treatment ($n = 1125$)[a] and at 1 year ($n = 1193$).

Outcomes	Post-treatment[a]	Missing, *n* %	1 year	Missing, *n* %
Pain intensity (*NRS, 1 to 10 scale*)[e], median [IQR]	2.0 [1.0 to 2.0]	591 (53)	2.0 [1.0 to 4.0]	552 (46)
Not recovered[b], *n %*	112 (21)		286 (45)	
Disability (*NDI, 0 to 50 scale*)[f], median [IQR]	5.0 [1.0 to 9.0]	628 (56)	5.0 [2.0 to 8.0]	515 (43)
Not recovered[c], n %	290 (58)		423 (62)	
Global perceived improvement (GPE, 7-point Likert scale)[g], *n %*		605 (54)		508 (43)
Completely recovered	127 (24)		149 (23)	
Much improved	287 (55)		247 (39)	
Slightly improved	83 (16)		143 (22)	
No change	24 (5)		81 (13)	
Slightly worse	0 (0)		11 (2)	
Much worse	0 (0)		8 (1)	
Worse than ever	0 (0)		2 (0)	
Not recovered[d], *n %*	107 (21)		264 (39)	

% rounded up to closest integer.
[a] Defined as no more than three months after intake, $n = 68$ not eligible.
[b] Not recovered >2, recovered ≤2.
[c] Score multiplied by 2 to yield %, not-recovered ≥8%, recovered <8%.
[d] Not recovered as “slightly improved”, “no change”, “slightly worse”, much worse”, “worse than ever”; recovered as “completely recovered” or “much improved”.
[e] NRS = numeric rating scale.
[f] NDI = neck disability index.
[g] GPE = global perceived effect.

Derived models

The derived models for post-treatment prediction are described in Table 3 and Supplement 1 and those for 1-year prediction in Table 4. The authors compared spline models’ performance to linear models’ performance for non-linear variable and outcome relations (i.e. Disability model at 1 year and Disability model post-treatment). Spline models’ performance appeared not superior to linear models’ performance and the authors choose to present these as linear models as they are more straightforward for clinical use. Models’ intercept, predictors, and assigned weights (beta’s) are displayed together with their performance and optimism-adjusted performance measures as evaluated in imputed data [8]. For all models the Hosmer–Lemeshow test was not-significant.

All derived post-treatment models exhibited acceptable discriminative performance. The disability model obtained the highest discriminative performance, and showed a calibration slope of 0.92 (IQR, 0.91–0.93), and R^2 of 0.24 (IQR, 0.22–0.26). The derived post-treatment pain and perceived improvement models exhibited somewhat lower discriminative performance, with calibration slope values of 0.86 (IQR, 0.91–0.93) and 0.86 (IQR, 0.84–0.87), respectively, and low explained variances. Calibration plots of post-treatment models are presented in Fig. 1. After adjustment for optimism, only the post-treatment disability model retained acceptable discriminative performance of AUC 0.74 (IQR, 0.72–0.75), and R^2 of 0.21 (IQR, 0.19–0.23).

None of the 1-year models reached the level of acceptable discriminative performance after derivation and after adjustment for optimism, and showed lower calibration slope values and explained variances.

Predictors in the models

Neck pain duration was a predictor in all models (Supplement 3). Baseline pain was a predictor in all pain models and baseline disability in all disability models. Age was a predictor included in all post-treatment models and headache in all 1-year models.

Sensitivity analyses

Sensitivity analyses on complete cases (post-treatment pain, disability, perceived improvement models, $n = 532$, 495, 518 respectively; 1-year pain, disability, perceived improvement models, $n = 476$, 508, 511 respectively) showed comparable performance measure values. The post-treatment pain model and the 1-year models derived in complete case data yielded the same or almost the same predictors (Supplement 3). The post-treatment disability model in the complete cases contained also sporting and previous episode as predictors and the perceived improvement model did not contain the sporting, previous episode, age and baseline disability predictors.

Table 3
Performance of prognostic models for predicting post-treatment recovery of neck pain ($n = 1193$)[#].

Predictors	Coefficient [##]	OR [##]	R^2	Optimism-adjusted R^2	AUC	Optimism-adjusted AUC
Pain model[a] *						
Constant	−3.62 (−4.66, −2.57)	0.03 (0.01 to 0.08)	0.13 [0.12 to 0.14] [\$]	0.09 [0.08 to 0.11] [\$]	0.70 (0.56 to 0.81) [\$\$]	0.67 [0.66 to 0.69] [\$]
Subacute pain	0.44 (−0.24, 1.13)	1.56 (0.78 to 3.10)				
Chronic pain	0.96 (0.47, 1.46)	2.62 (1.60 to 4.31)				
Baseline pain (NRS 0 to 10)[d]	0.19 (0.07, 0.31)	1.21 (1.07 to 1.36)				
BNQ anxiety & depression (0 to 20)[e]	0.04 (−0.00, −0.10)	1.05 (1.00 to 1.10)				
Age	0.01 (−0.00, 0.02)	1.01 (1.00 to 1.02)				
Disability model[b] **						
Constant	−2.75 (−3.58, −1.93)	0.06 (0.03 to 0.15)	0.24 [0.22 to 0.26] [\$]	0.21 [0.19 to 0.23] [\$]	0.75 (0.63 to 0.84) [\$\$]	0.74 [0.72 to 0.75] [\$]
Subacute pain	0.30 (−0.27, 0.86)	1.34 (0.77 to 2.36)				
Chronic pain	0.96 (0.53, 1.40)	2.62 (1.70 to 4.03)				
Baseline disability (NDI 0 to 50[f]	0.12 (0.08, 0.16)	1.13 (1.08 to 1.17)				
Age	0.02 (0.01, 0.03)	1.02 (1.01 to 1.03)				
General sleeping problems	0.31 (−0.10, 0.72)	1.36 (0.91 to 2.05				
FABQ physical activity (0 to 24)[g]	0.02 (−0.01, 0.05)	1.02 (0.99 to 1.05)				
Perceived improvement model[c] ***						
Constant	−2.72 (−3.80, −1.64)	0.07 (0.02 to 0.19)	0.13 [0.11 to 0.15] [\$]	0.09 [0.07 to 0.11] [\$]	0.70 (0.56 to 0.80) [\$\$]	0.67 [0.65 to 0.69] [\$]
Subacute pain	0.16 (−0.70, 1.03)	1.18 (0.49 to 2.81)				
Chronic pain	0.95 (0.46, 1.43)	2.57 (1.60 to 4.17)				
Low back pain	0.41 (−0.02, 0.84)	1.51 (0.98 to 2.30)				
FABQ physical activity (0 to 24)	0.04 (0.00, 0.08)	1.04 (1.00 to 1.08)				
Age	0.01 (−0.00, 0.03)	1.02 (0.10 to 1.03)				
Baseline disability (NDI 0 to 50)	−0.03 (−0.07, 0.01)	0.97 (0.93 to 1.01)				
Previous episode	−0.46 (0.00, 0.08)	1.04 (1.00 to 1.08)				
Partaking in sporting activities	0.38 (−0.05, 0.81)	1.46 (0.95 to 2.25)				

[#]Imputed data; [##] In logit scale as mean with 95% confidence interval (CI); [\$] In logit scale as median with interquartile range [IQR]; [\$\$] In logit scale as mean with 95% CI.

[a] Pain intensity measured with NRS *1–10-point Likert scale);* not-recovered >2.

[b] Disability measured with NDI [2] (*0–50 scale*, sum score *multiplied by 2 to yield %);* not-recovered ≥8%.

[c] General Perceived Effect measured with GPE [3] (*7-point Likert scale);* non-recovered as “slightly improved”, “no change”, “slightly worse”, much worse”, “worse than ever”.

[d] NRS = numeric rating scale (*1–10-point Likert scale*).

[e] NBQ-AD = Neck Bournemouth Questionnaire, anxiety and depression subscale (*scale 0–20*), sum score of 11-point numeric subscale of items 4 and 5.

[f] NDI = neck disability index (*0–50 scale*).

[g] FABQ-PA = Fear Avoidance Beliefs Questionnaire, Physical Activity subscale (*scale 0–24*).

Table 4

Performance of prognostic models for predicting 1-year recovery of neck pain ($n = 1193$).#.

Predictors	Coefficient ##	OR ##	R^2	Optimism-adjusted R^2	AUC	Optimism-adjusted AUC
Pain model[a]						
Constant	−1.27 (−1.75, −0.80)	0.28 (0.17 to 0.45)				
Baseline pain (NRS 0 to 10)[d]	0.13 (0.06, 0.21)	1.14 (1.06 to 1.24)				
General sleeping problems	−0.48 (−0.85, −0.12)	0.62 (0.43 to 0.88)	0.09 [0.08 to 0.10] $	0.06 [0.05 to 0.07] $	0.65 (0.52 to 0.76) $$	0.62 [0.62 to 0.63] $
Previous episode	0.29 (−0.04, 0.62)	1.34 (0.96 to 1.86)				
Low back pain	0.33 (0.03, 0.64)	1.40 (1.03 to 1.89)				
Headache	0.30 (0.00, −0.60)	1.35 (1.00 to 1.83)				
Disability model[b]						
Constant	−1.01 (−1.69, 0.33)	0.36 (0.18 to 0.71)				
Subacute pain	0.05 (−0.4, 0.51)	1.05 (0.66 to 1.67)				
Chronic pain	0.48 (0.13, 0.84)	1.62 (1.13 to 2.32)	0.09 [0.08 to 0.10] $	0.06 [0.05 to 0.07] $	0.65 (0.53 to 0.76) $$	0.63 [0.62 to 0.64] $
Baseline disability (NDI 0 to 50)[e]	0.05 (0.02, 0.08)	1.05 (1.02 to 1.08)				
Age	0.01 (0.00, 0.02)	1.01 (1.00 to 1.02)				
Headache	0.36 (0.01, 0.72)	1.44 (1.01 to 2.05)				
Perceived improvement model[c]						
Constant	−1.38 (−1.85, −0.92)	0.25 (0.16 to 0.40)				
Subacute pain	0.37 (−0.16, 0.91)	1.45 (0.85 to 2.49)				
Chronic pain Baseline disability (NDI 0 to 50)	0.40 (0.03, 0.77) 0.04 (0.01, 0.06)	1.49 (1.03 to 2.15) 1.04 (1.01 to 1.06)	0.10 [0.09 to 0.11] $	0.07 [0.06 to 0.08] $	0.66 (0.53 to 0.77) $$	0.64 [0.63 to 0.65] $
Low back pain	0.46 (0.13, 0.79)	1.58 (1.13 to 2.20)				
General sleeping problems	−0.40 (−0.76, −0.03)	0.67 (0.47 to 0.97)				
Female gender	−0.37 (−0.73, −0.01)	0.70 (0.48 to 0.99)				
Headache	0.54 (0.22, 0.86)	1.72 (1.25 to 2.38)				

#Imputed data; ## In logit scale as mean with 95% confidence interval (CI); $ In logit scale as median with interquartile range [IQR]; $$ In logit scale as mean with 95% CI.

[a] Pain intensity measured with NRS *1–10-point Likert scale);* not-recovered >2.

[b] Disability measured with NDI [2] (*0–50 scale*, sum score *multiplied by 2 to yield %);* not-recovered ≥8%.

[c] General Perceived Effect measured with GPE [3] (*7-point Likert scale);* non-recovered as “slightly improved”, “no change”, “slightly worse”, much worse”, “worse than ever”.

[d] NRS = numeric rating scale (*1–10-point Likert scale*).

[e] NDI = neck disability index (*0–50 scale*).

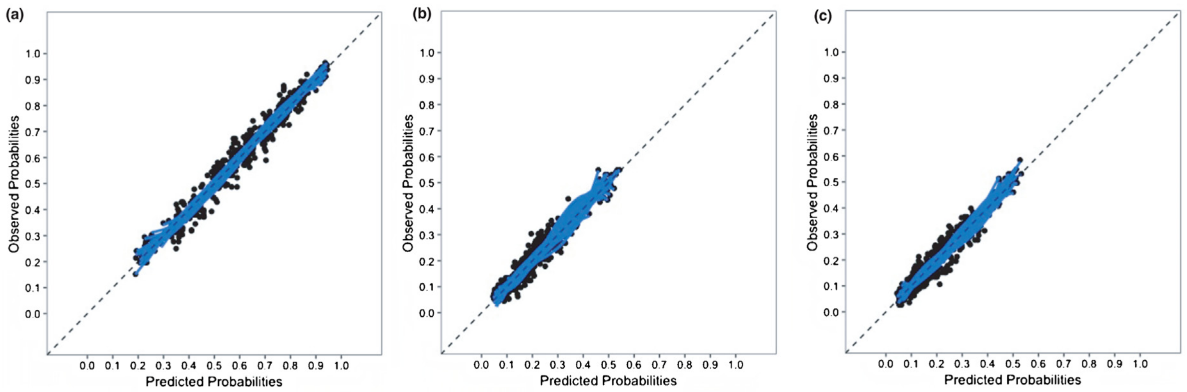

Fig. 1. Calibration plots. a. Disability model. b. Pain model. c. Perceived improvement model.

Discussion

Main result

The derived model for post-treatment disability containing baseline pain duration, baseline disability, age, sleeping problem and FABQ-physical activity as predictors exhibited the best overall performance, calibration, and discrimination and it also exceeded the threshold for acceptable discriminative ability after adjustment for optimism. The other post-treatment models almost reached acceptable discriminative ability after adjustment. None of the derived 1-year models reached acceptable discriminative performance and showed lower calibration slope values and explained variances.

Important results models

The post-treatment models performed better than the 1-year models and exhibited discrimination of 0.70 or upward and calibration slopes more or less around a value of 0.90. Is seems plausible that short-term prediction is more accurate compared to long-term prediction. The post-treatment disability model performed best, possibly because the outcome was measured with the NDI, which is an instrument that covers various health constructs [52]. The NRS is a single-item questionnaire which measures a narrower domain and may also have larger measurement error that can influence the performance of the models [53]. The same may apply to the GPE which, additionally, is an instrument reflecting the current health status more than change in health status over time [36].

On the whole, our derived models, especially the post-treatment disability model, performed better as compared to existing models that predict recovery in neck pain patients, although few derivation studies allow proper comparison of model performance as both discrimination and calibration performance measures were seldom presented [15,54].

Important results predictors

Neck pain duration was a predictor in all models and independent of type of outcome or follow-up time. Baseline disability was a predictor in almost all models except for pain outcome. Baseline pain was a predictor in almost all models except for disability outcome. Age was a predictor that corresponded consistently with post-treatment follow-up and headache with 1-year follow-up.

Model comparison with literature

One study with six months follow up and a GPE outcome derived a model in a primary care population ($n = 468$) treated for non-serious neck pain and validated this model in a primary care setting treated with manual therapy and electrotherapy ($n = 346$) [35]. This model performed less well if compared to the post-treatment model on GPE outcome in our study but similarly to the 1-year model. Its external validation study revealed a possibly helpful discriminative ability of AUC 0.65 (95% CI, 0.59 to 0.71), a value slightly better compared to our internal validation [35]. Another study developed models also using the NDI as an outcome in people with acute whiplash associated disorder (WAD) at one-year [55]. Models' overall performance (R^2) was presented but no model calibration and discrimination were calculated, which hampers comparison of model performance. However, these models performed not well at external validation [16]. Another study developed a prognostic model for WAD, with six months follow-up, in an insurance company subcohort treated with physical therapy physiotherapy and collected self-reported recovery outcome through telephone interview [56]. An AUC of 0.67 (95% CI, 0.63 to 0.70) was reported after internal validation. This is comparable to the post-treatment model on GPE outcome after internal validation in our study and somewhat better compared to our 1-year model after internal validation. In the current study, the authors recruited patients with non-specific neck pain of any duration including neck pain with trauma, and, in contrast with the two aforementioned studies, the authors did not develop a model specific for WAD.

Predictors in the models compared with literature

A recent overview of systematic reviews on prognostic factors in neck pain reported that higher baseline NDI and pain at inception were predictors of outcomes after WAD [11]. In our study, in which patients with non-specific neck pain and WAD were included, all models that predicted disability yielded baseline disability as predictor, and models that predicted pain contained baseline pain as predictor. This is in line with the vast majority of models that predicted disability outcomes and pain outcomes as described in a recent systemic review [15]. Baseline NDI and baseline pain are consistent reported prognostic factors [11,24,25] for prediction of disability and neck pain, respectively. This is also the case for neck pain duration as a consistently reported prognostic factor [11,24,25] that retains its predictive ability in relation to other prognostic factors for all outcomes as well as age and headache who are consistently reported prognostic factors [11,24,25] that retain their predictive ability in relation to other prognostic factors, for short-term and long-term prognosis, respectively. Sex and previous neck pain episode [11,24,25] appeared less consistent in relation to other prognostic factors.

Strengths and limitations

In contrast with previously published prognostic models for neck pain [15] the models in our study were developed in a large cohort with sufficient power, and the cohort closely resembles clinical practice in primary care manual therapy in The Netherlands. The authors used the most recent methods in terms of *a priori* model sample size calculation, development, and internal validation. After internal validation, the authors presented penalized full models for the models that demonstrated acceptable performance.

The main limitation of this study is the cohort's missing data, especially for the outcome variables. The high dropout can be explained by the fact that participants returned outcome questionnaire booklets by post that had to be number marked by themselves and when this was missing the booklets could be labelled at their arrival by the researchers. However, the labels with the patient number on them were frequently lost or separated from the booklets and then the total questionnaire information could not be used anymore. The authors think due to these reasons that the underlying missing data mechanism tends towards an MCAR and MAR mechanism but certainly not MNAR. Also, because the majority of predictors are shared by MI and complete case data, especially for one-year follow-up and baseline characteristics were comparable between complete cases and those without outcome data. As recommended in the literature [41], missing value analysis was conducted and multivariable multiple imputation on predictors as well as outcomes with an amount of 50 imputed sets. There is some evidence from simulation studies that this high missing data rate can be handled with multiple imputation [57]. To address the potential limitation of gain from imputation, complete case analyses were also performed as sensitivity analyses and these showed very similar parameter estimates and this consistency supports our conclusions. Another limitation to be addressed is that the authors used binary outcomes for reason of comparison with previous developed models. The use of other cut-offs may have resulted in other model predictors or model performance and the derived models have to be interpreted in relation to the cut-offs points used at issue.

The authors reached sample size for the post-treatment disability model and all 1-year models. However, the post-treatment pain and perceived improvement models fell one predictor parameter short to reach effective sample size (the excluded candidate factor for treatment expectations was considered). The authors believe to have corrected for this overfitting by penalizing the post-treatment models after internal validation.

Conclusions

A post-treatment prognostic model for disability was successfully developed and internally validated. This model has potential to inform primary care clinicians about a patient's individual prognosis after treatment, but external validation is required before broad clinical use can be recommended.

Implications for practice and further research

Recovery is a multidimensional construct and clinical guidelines usually promote the use of several outcome measures simultaneously [58]. For this reason, the authors propose that, if all adequately performing during external validation, the future potential clinical use will be of all the three separate models developed in this study. The post-treatment models for prediction of recovery in patients with non-specific neck pain, especially the disability model, have good potential for clinical use. The post-treatment disability model can inform clinicians at intake about patient's individual prognosis after therapy. To illustrate this for an intake situation where a physiotherapist wants to inform a neck pain patient about his or her specific prognosis: "based on this model and you being 30 years of age, having 10 weeks neck pain duration, a 7/50 NDI score, sleeping problems and a 4/24 FABQ-PA score, the authors expect there is a 35% chance you will not be recovered post-treatment (or *vice-versa* a 65% chance that you will be recovered after treatment). However, before clinical use can be promoted, the authors suggest post-treatment models' further external validation, especially the disability model. The post-treatment disability model derived in our study showed precise optimism-adjusted AUC of 0.74 with small 95% CI width of 0.03. The authors argue this is a promising value for external validation, given our pursuit to avoid key methodological shortcomings and therefore likely obtaining models that are less overfitted than the large majority of those developed for neck pain so far [15]. Additionally, the post-treatment pain and perceived improvement models exhibited also precise optimism-adjusted AUCs of 0.67 with small 95% CI widths of 0.03 and 0.04, respectively. The authors strongly believe there is room to expand models' performance by updating these models with other predictors that were not evaluated in the ANIMO cohort (e.g., clinical examination findings).

The models' relatively low explained variances indicate potential for improvement with relevant predictors that are still missing and literature knowledge seems to provide us only limited information. Further research on new predictors that can strengthen the models is needed. Furthermore, the authors suggest research on predictors of treatment effect (e.g. by randomized controlled trials), since they could not be accounted for in this single cohort study design. Specifically, causally related modifiable factors have potential to change patient outcome [8].

Ethics approval: Obtained from Erasmus Medical Centre, Rotterdam, the Netherlands (Ref. No. MEC-2007-359).

Data was deidentified and the study complies with the Declaration of Helsinki.

Funding: This study was partly funded by a program grant of the Dutch Arthritis Foundation. The funder played no role in the design, conduct, or reporting of this study.

Conflict of interest: None declared.

Availability of data: The dataset used and analyzed during the current study is available through SOMT upon reasonable request.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.physio.2021.05.011.

References

[1] Abajobir AA, Abate KH, Abbafati C, Abbas KM, Abd-Allah F, Abdulkader RS, *et al.* Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017;390:1260–344, http://dx.doi.org/10.1016/S0140-6736(17)32130-X.

[2] 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;92:824–9, http://dx.doi.org/10.1016/j.apmr.2010.12.025.

[3] Vasseljen O, Woodhouse A, Bjørngaard JH, Leivseth L. Natural course of acute neck and low back pain in the general population: the HUNT study. Pain 2013;154:1237–44, http://dx.doi.org/10.1016/j.pain.2013.03.032.

[4] Hurwitz EL, Randhawa K, Yu H, Côté P, Haldeman S. The global spine care initiative: a summary of the global burden of low back and neck pain studies. Eur Spine J 2018:1–6, http://dx.doi.org/10.1007/s00586-017-5432-9.

[5] Babatunde OO, Jordan JL, Van der Windt DA, Hill JC, Foster NE, Protheroe J. Effective treatment options for musculoskeletal pain in primary care: a systematic overview of current evidence. PLoS One 2017;12:e0178621, http://dx.doi.org/10.1371/journal.pone.0178621.

[6] Coulter ID, Crawford C, Vernon H, Hurwitz EL, Khorsan R, Booth MS, *et al.* Manipulation and mobilization for treating chronic nonspecific neck pain: a systematic review and meta-analysis for an appropriateness panel. Pain Physician 2019;22:E55–70.

[7] Gross A, Langevin P, Burnie SJ, Bédard-Brochu M-S, Empey B, Dugas E, *et al.* Manipulation and mobilisation for neck pain contrasted against an inactive control or another active treatment. Cochrane Database Syst Rev 2015;(9):CD004249, http://dx.doi.org/10.1002/14651858.CD004249.pub4.

[8] Riley RD, van der Windt DA, Croft P, Moons KGM. Prognosis research in health care, concepts, methods, and impact. 1st ed. Oxford: Oxford University Press; 2019.

[9] Artus M, Campbell P, Mallen CD, Dunn KM, van der Windt DAW. Generic prognostic factors for musculoskeletal pain in primary care: a systematic review. BMJ Open 2017;7:e012901, http://dx.doi.org/10.1136/bmjopen-2016-012901.

[10] Carroll LJ, Hogg-Johnson S, van der Velde G, Haldeman S, Holm LW, Carragee EJ, *et al.* Course and prognostic factors for neck pain in the general population. Spine (Phila Pa 1976) 2008;33:S75–82 https://doi.org/10.1097/BRS.0b013e31816445be.

[11] Walton DM, Carroll LJ, Kasch H, Sterling M, Verhagen AP, Macdermid JC, *et al.* An overview of systematic reviews on prognostic factors in neck pain: results from the International Collaboration on Neck Pain (ICON) project. Open Orthop J 2013;7:494–505 https://doi.org/10.2174/1874325001307010494.

[12] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338, http://dx.doi.org/10.1136/bmj.b375, b375–b375.

[13] Bordini BJ, Stephany A, Kliegman R. Overcoming diagnostic errors in medical practice. J Pediatr 2017;185:19–25.e1, http://dx.doi.org/10.1016/j.jpeds.2017.02.065.

[14] Chiffi D, Zanotti R. Fear of knowledge: clinical hypotheses in diagnostic and prognostic reasoning. J Eval Clin Pract 2016:1–7, http://dx.doi.org/10.1111/jep.12664.

[15] 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;64:16–23, http://dx.doi.org/10.1016/j.jphys.2017.11.013.

[16] Wingbermühle RW, Heymans MW, Trijffel E van, Koes B, Arianne P. Verhagen. External validation study of three promising models for prediction of neck pain recovery. Submitted n.d.

[17] Peters R, Mutsaers B, Verhagen AP, Koes BW, Pool-Goudzwaard AL. Prospective cohort study of patients with neck pain in a manual therapy setting: design and baseline measures. J Manipulative Physiol Ther 2019;42:471–9, http://dx.doi.org/10.1016/j.jmpt.2019.07.001.

[18] Steyerberg E, Moons KGM, van der Windt D, Hayden J, Perel P, Schroter S, *et al.* Prognosis research strategy (PROGRESS) series 3: prognostic models. Br Med J 2012;10, http://dx.doi.org/10.1371/jour-.

[19] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55, http://dx.doi.org/10.7326/M14-0697.

[20] Heinze G, Wallisch C, Dunkler D. Variable selection—a review and recommendations for the practicing statistician. Biom J 2018:1–19, http://dx.doi.org/10.1002/bimj.201700067.

[21] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, *et al.* PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1, http://dx.doi.org/10.7326/M18-1377.

[22] Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009;338:1373–7, http://dx.doi.org/10.1136/bmj.b604.

[23] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31, http://dx.doi.org/10.1093/eurheartj/ehu207.

[24] Bruls VEJ, Bastiaenen CHG, de Bie RA. Prognostic factors of complaints of arm, neck, and/or shoulder. Pain 2015;156:765–88, http://dx.doi.org/10.1097/j.pain.0000000000000117.

[25] Carroll LJ, Hogg-Johnson S, Van Der Velde G, Haldeman S, Holm LW, Carragee EJ, *et al.* Course and prognostic factors for neck pain in the general population: results of the bone and joint decade 2000–2010 Task Force on Neck Pain and its associated disorders. Spine (Phila Pa 1976) 2008;33:S75–82.

[26] Landers MR, Creger RV, Baker CV, Stutelberg KS. The use of fear-avoidance beliefs and nonorganic signs in predicting prolonged disability in patients with neck pain. Man Ther 2008;13:239–48, http://dx.doi.org/10.1016/j.math.2007.01.010.

[27] Lundberg M, Grimby-Ekman A, Verbunt J, Simmonds MJ. Pain-related fear: a critical review of the related measures. Pain Res Treat 2011;2011, http://dx.doi.org/10.1155/2011/494196.

[28] Geri T, Piscitelli D, Meroni R, Bonetti F, Giovannico G, Traversi R, *et al.* Rasch analysis of the Neck Bournemouth Questionnaire to measure disability related to chronic neck pain. J Rehabil Med 2015;47:836–43, http://dx.doi.org/10.2340/16501977-2001.

[29] Geri T, Signori A, Gianola S, Rossettini G, Grenat G, Checchia G, *et al.* Cross-cultural adaptation and validation of the Neck Bournemouth Questionnaire in the Italian population. Qual Life Res 2015;24:735–45, http://dx.doi.org/10.1007/s11136-014-0806-5.

[30] Schmitt MA, de Wijer A, van Genderen FR, van der Graaf Y, Helders PJ, van Meeteren NL. The Neck Bournemouth Questionnaire cross-cultural adaptation into Dutch and evaluation of its psychometric properties in a population with subacute and chronic whiplash associated disorders. Spine (Phila Pa 1976) 2009;34:2551–61, http://dx.doi.org/10.1097/BRS.0b013e3181b318c4.

[31] Palmlöf L, Holm LW, Alfredsson L, Skillgate E. Expectations of recovery: a prognostic factor in patients with neck pain undergoing manual therapy treatment. Eur J Pain 2016;20:1384–91, http://dx.doi.org/10.1002/ejp.861.

[32] Wright Aa, Hensley Cp, Gilbertson J, Leland Jm, Jackson S. Defining patient acceptable symptom state thresholds for commonly used patient reported outcomes measures in general orthopedic practice. Man Ther 2015;20:814–9, http://dx.doi.org/10.1016/j.math.2015.03.011.

[33] Sterling M, Jull G, Kenardy J. Physical and psychological factors maintain long-term predictive capacity post-whiplash injury. Pain 2006;122:102–8.

[34] MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, *et al.* Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther 2009;39:400–17, http://dx.doi.org/10.2519/jospt.2009.2930.

[35] Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HCW, Koes BW. Prognosis of patients with nonspecific neck pain. Spine (Phila Pa 1976) 2010;35:E827–835, http://dx.doi.org/10.1097/BRS.0b013e3181d85ad5.

[36] Kamper SJ, Ostelo RWJG, Knol DL, Maher CG, de Vet HCW, Hancock MJ. Global perceived effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. J Clin Epidemiol 2010;63:760–6.e1, http://dx.doi.org/10.1016/j.jclinepi.2009.09.009.

[37] Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM, Van Der Heijden G. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087–91, http://dx.doi.org/10.1016/j.jclinepi.2006.01.014.

[38] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77, http://dx.doi.org/10.1037/1082-989X.7.2.147.

[39] Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol 2010;63:205–14, http://dx.doi.org/10.1016/j.jclinepi.2009.03.017.

[40] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:1–10, http://dx.doi.org/10.1136/bmj.b2393.

[41] Lee KJ, Roberts G, Doyle LW, Anderson PJ, Carlin JB. Multiple imputation for missing data in a longitudinal cohort study: a tutorial based on a detailed case study involving imputation of missing outcome data. Int J Soc Res Methodol 2016;19:575–91, http://dx.doi.org/10.1080/13645579.2015.1126486.

[42] Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol 2009;9:1–8, http://dx.doi.org/10.1186/1471-2288-9-57.

[43] Heymans MW. R package psfmi: Predictor Selection Functions for Logistic and Cox regression models in multiply imputed datasets; 2019, 0.1.0 2019.

[44] Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ 2006;332:1080, http://dx.doi.org/10.1136/bmj.332.7549.1080.

[45] Collins GS, Ogundimu EO, Cook JA, Le Manach Y, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med 2016;35:4124–35, http://dx.doi.org/10.1002/sim.6986.

[46] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127–41, http://dx.doi.org/10.1002/sim.2331.

[47] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, 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:W1–73, http://dx.doi.org/10.7326/M14-0698.

[48] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Wiley; 2013.

[49] Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81, http://dx.doi.org/10.1016/S0895-4356(01)00341-9.

[50] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76, http://dx.doi.org/10.1016/j.jclinepi.2015.12.005.

[51] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell Jr FE, Moons KGM, *et al.* Minimum sample size for developing a multivariable prediction model: PART II—binary and time-to-event outcomes. Stat Med 2019;38:1276–96, http://dx.doi.org/10.1002/sim.7992.

[52] Ailliet L, Knol DL, Rubinstein SM, De Vet HCW, Van Tulder MW, Terwee CB. Definition of the construct to be measured is a prerequisite for the assessment of validity. The neck disability index as an example. J Clin Epidemiol 2013;66:775–82.e2, http://dx.doi.org/10.1016/j.jclinepi.2013.02.005.

[53] Chiarotto A, Maxwell LJ, Ostelo RW, Boers M, Tugwell P, Terwee CB. Measurement properties of visual analogue scale, numeric rating scale, and pain severity subscale of the brief pain inventory in patients with low back pain: a systematic review. J Pain 2019;20:245–63, http://dx.doi.org/10.1016/j.jpain.2018.07.009.

[54] Kelly J, Ritchie C, Sterling M. Clinical prediction rules for prognosis and treatment prescription in neck pain: a sys-

tematic review. Musculoskelet Sci Pract 2017;27:155–64, http://dx.doi.org/10.1016/j.math.2016.10.066.

[55] Ritchie C, Hendrikz J, Kenardy J, Sterling M. Derivation of a clinical prediction rule to identify both chronic moderate/severe disability and full recovery following whiplash injury. Pain 2013;154:2198–206, http://dx.doi.org/10.1016/j.pain.2013.07.001.

[56] Bohman T, Cote P, Boyle E, Cassidy JD, Carroll LJ, Skillgate E. Prognosis of patients with whiplash-associated disorders consulting physiotherapy: development of a predictive model for recovery. BMC Musculoskelet Disord 2012;13:264, http://dx.doi.org/10.1186/1471-2474-13-264.

[57] Kontopantelis E, White IR, Sperrin M, Buchan I. Outcome-sensitive multiple imputation: a simulation study. BMC Med Res Methodol 2017;17:1–13, http://dx.doi.org/10.1186/s12874-016-0281-5.

[58] 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 Care Res (Hoboken) 2008;61:124–31, http://dx.doi.org/10.1002/art.24162.

Available online at www.sciencedirect.com

ScienceDirect