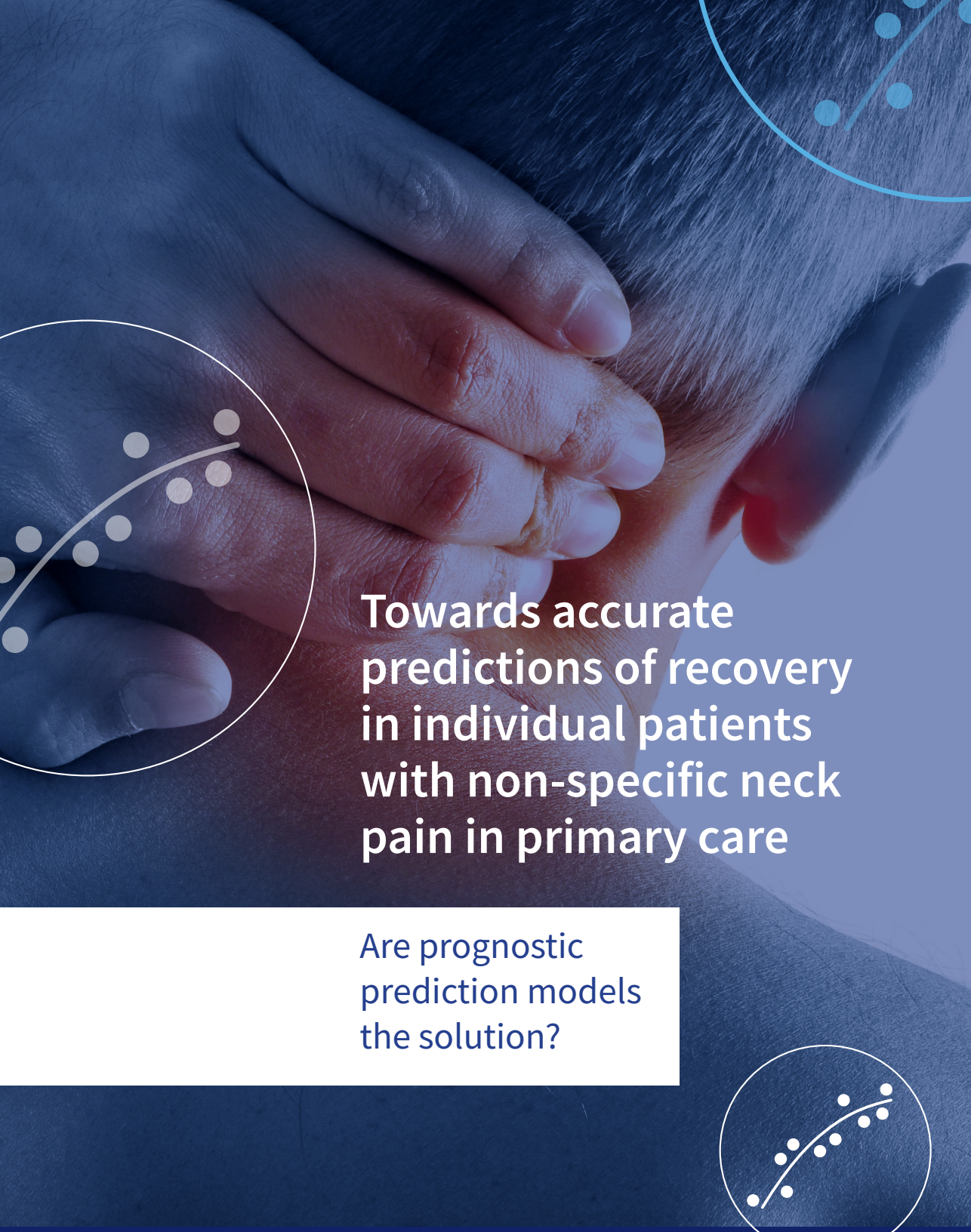


Towards accurate predictions of recovery in individual patients with non-specific neck pain in primary care

Are prognostic prediction models the solution?

Roel W. Wingbermühle

Towards accurate predictions of recovery in individual patients with non-specific neck pain in primary care

Are prognostic prediction models the solution?

Roel W. Wingbermühle

Towards Accurate Predictions of Recovery in Individual Patients with Non-specific Neck Pain in Primary Care

Are prognostic prediction models the solution?

Naar accurate voorspellingen van het herstel bij individuele patiënten met niet-specifieke nekpijn in de eerste lijnsgezondheidszorg

Zijn prognostische voorspelmodellen de oplossing?

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof. dr. A.L. Bredenoord

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

donderdag 15 december 2022 om 15.30 uur

door

Roeland Willem Wingbermühle
geboren te Amersfoort.

Erasmus University Rotterdam

Colophon
PhD thesis, Erasmus Universiteit Rotterdam, The Netherlands

Printed by: Graphic in Mind
Design: BuroPARK

ISBN: 978-94-6437-718-7

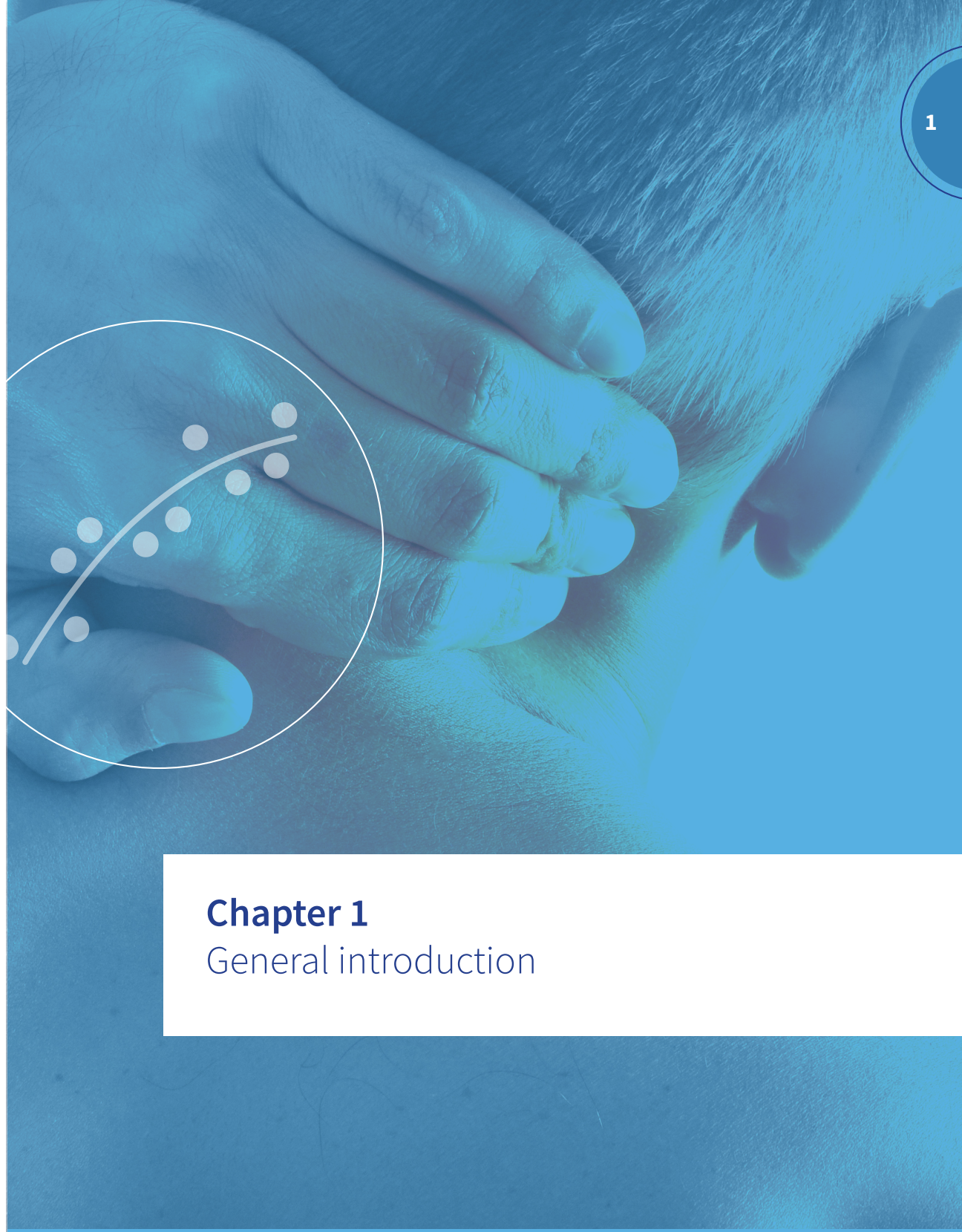
somt
University of Physiotherapy

Financial support for printing and distribution of the thesis was kindly provided by Erasmus MC and SOMT University of Physiotherapy.

PROMOTIECOMMISSIE

Promotor:	Prof. dr. B.W. Koes
Overige leden:	Prof. dr. ir. A. Burdorf Prof. dr. C. Lucas Prof. dr. B. Cagnie
Copromotor:	Dr. A. Chiarotto

Table of contents



Chapter 1
General introduction

Chapter 1. General introduction

The burden of neck pain

Neck pain is a common global health problem leading to substantial pain, disability, and economic costs in most countries. [1, 2, 3] The burden of neck pain was demonstrated in the 2019 Global Burden of Disease (GBD) study by an all-age global prevalence of 223 million people (179-281) and 22 million (15-31) Years of life Lived with Disability (YLD), showing a need for rehabilitation of this condition. [1] A specific analysis of the burden of neck pain from the 2017 GBD study revealed that the global number of prevalent neck pain cases increased with age to peak at middle age (45-49 years for men, 45-54 years for women), and then decreased with older age, where the number of cases was higher in females across all age groups. [3] In terms of global causes of YLDs, neck pain ranked 9th out of 354 conditions for females and 11th for males in 2017. [4] Extrapolating the 2019 neck pain incidence and prevalence numbers in general practice in the Netherlands, as registered per 1000 persons by the Dutch institute for health research (NIVEL), showed that the neck pain incidence in women was 492.456 and in men 310.386, and the neck pain prevalence in women was 443.904 and in men, 284.376. [5] The cervical-thoracic diagnostic code was the most common diagnostic code registered by physiotherapists in the Netherlands between 2014 and 2017. [6 7] From 2018 to 2020 it was the second most common registered code, after the lumbar diagnostic code. [7] This indicates a high incidence of people with neck pain problems in physiotherapy practice in the Netherlands.

How to classify patients with neck pain?

Neck pain definition may vary, but it is most often defined as pain in the cervical spine region with or without referred pain to the arms. [3 8] This is often time-based subdivided into acute (0-6 weeks), subacute (6-12 weeks) or chronic neck pain (>3 months). [9 10] However, guidelines differ in their definitions of acute and sub acute time frames. [11]
In the vast majority of patients no pathoanatomical cause can be identified, thereby the neck pain is labelled as non-specific neck pain. [12] A four-grade classification system was developed by the Neck Pain Task Force, based on the amount of interference with daily living activities, signs and symptoms suggestive of structural pathology, and the presence (use an article) of neurological signs (see Table 1). [13 14] Also, neck pain can be associated with a work-related condition, or a traumatic injury such as a motor vehicle accident or whiplash associated disorder, which can be labelled as traumatic neck pain versus non-traumatic neck pain. [15] Some classification systems do not differentiate between traumatic and non-traumatic neck pain and some research shows that there are no clinically relevant differences between them regarding pain, function, or prognosis. [13 16] Other researchers argue that there are clinically relevant differences between traumatic and non-traumatic neck pain, and it is suggested that the group with traumatic neck pain is a relevant subgroup with a worse prognosis that needs different treatment. [15,17]

Grade	Explanation
I	Neck pain and associated disorders with no signs or symptoms suggestive of major structural pathology and no or minor interference with activities of daily living
II	No signs or symptoms of major structural pathology, but major interference with activities of daily living
III	No signs or symptoms of major structural pathology, but presence of neurologic signs such as decreased deep tendon reflexes, weakness, or sensory deficits in the upper extremity
IV	Signs or symptoms of major structural pathology, which include (but are not limited to) fracture, vertebral dislocation, injury to the spinal cord, infection, neoplasm, or systemic disease including inflammatory arthropathies

Table 1. Classification system by the Neck Pain Task Force [13 14]

Course and prognosis of neck pain

Recent studies showed that recovery of neck pain mainly takes place in the first 4-6 weeks, without further evident reduction of neck pain and disability afterwards. [18 19] The Neck Pain Task Force indicated in 2008 that between 50%-85% of people who experience an episode of neck pain will report neck pain again within 1-5 years. [20] They conclude that the course of people with general neck pain, traumatic neck pain and work-related neck pain is remarkably similar, and the prognosis of neck pain is multi-factorial. [21 22]
In general practice in the Netherlands, 47% of acute non-specific neck pain patients reported having neck pain still at a 1-year follow-up, and 5.6% reported a recurrence. [10]
This indicates that, when people do not recover within the first few weeks, the prognosis for a group of people leads to persistent or intermittent pain and disability. Therefore, identification of patients very likely to recover in the short-term may reduce the risk of overtreatment and health costs by providing patients reassurance and self-management advice, instead of a “treat-all policy”. However, early identification of neck pain patients with expected worse outcomes enables clinicians to offer effective treatments that may abate patients' burden and health costs. Besides reassurance, advice and education, the current recommended treatments for neck pain in general are exercise therapy, analgesics, and manual therapy combined with other modalities. [23 11] Psychological, multimodal, or multidisciplinary interventions are recommended for subgroups of patients with psychosocial risk factors or chronic neck pain. [23] For chronic neck pain, available evidence shows the strongest effect for exercise and small effects for advice, education and psychological treatment, however further research is necessary and may likely change the effect estimates. [15]

Why focus on prognostic research?

Patients with neck pain have concerns about their future and one of the most important aspects to know for patients consulting a clinician is the likely prognosis of their condition. [24] In this regard, prognosis concerns the expected future outcome of an individual patient's health condition which can guide shared clinical treatment decisions and lifestyle changes. [25] For many health conditions, including musculoskeletal health conditions such as neck pain and low back pain, observational studies provide information on the average course or outcome of that health condition. Recovery from neck pain and neck pain related disability mainly takes place in the first few weeks without further subsequent improvement.
In patients with acute neck pain, the pooled weighted mean neck pain score of 64 (95% CI, 61-67, 0-100 scale) at onset decreases to 35 (95% CI, 32-38) at 6.5 weeks and 42 (95% CI, 39-45) at 12 months (see Figure 1). [18] The pooled weighted mean disability score of 30 (95% CI, 28-32) at onset decreases to 17 (95% CI, 15-19) at 6.5 weeks without further improvement at 12 months (see Figure 2). [18]

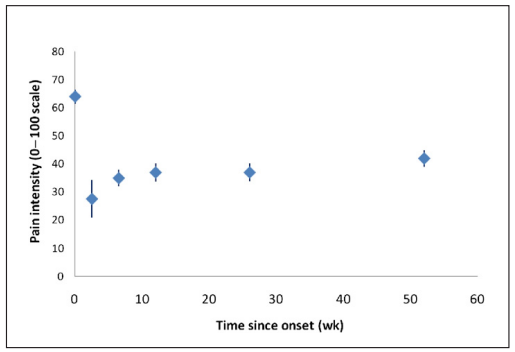


Fig 1. Course of neck pain intensity*

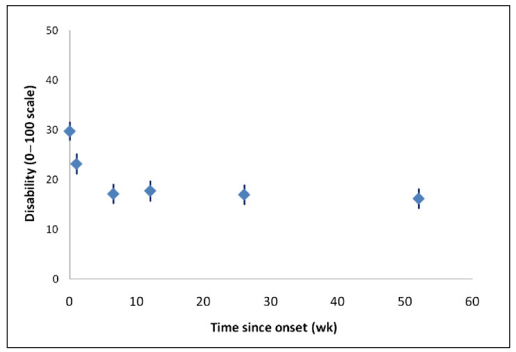


Fig 2. Course of neck pain disability*

***Pooled weighted means and 95% CI. Endpoints are combined from multiple studies and studies with multiple endpoints[18]**

However, a patient is usually interested in his or her individual prognosis, which is very likely a deviation from the average course or outcome, based on the patient's individual characteristics. Informing an individual neck pain patient about their prognosis is a challenging task for a primary care clinician, including myself. [26] In treating patients with spinal health problems such as neck pain, I still experience difficulties in answering patients' prognostic questions like: "will I recover from my neck pain?" or "how long will it take to recover from my neck pain?" The challenge I experienced in answering my patients' prognostic questions, inspired me to conduct this thesis. A clinician may estimate a patient's individual outcome which provides a limited accurate prognosis. [27 28] Prognostic factors and prognostic prediction models (further prognostic models) can provide a clinician with additional information to improve the estimation of the patients' individual prognosis and subsequent clinical decisions.

A prognostic factor is a variable that, in a given health condition, is associated with a future outcome. [29] Prognosis can incorporate a wide range of information and prognostic factors can be based on e.g., demographic factors, disease characteristics, history taking, physical examination, treatment, or additional examinations such as imaging, blood assays, urine tests or other biological measurements. [30] A recent review of systematic reviews judged seven prognostic factors for predicting future commonly used outcomes in neck pain and back pain as having moderate confidence for robust findings (i.e., disability/activity limitations, mental health, pain intensity, pain severity, coping, expectations of outcome/recovery, and fear avoidance). [31] Prognostic models combine values of multiple prognostic factors and have the advantage of estimating future health outcomes of individual patients. [29 32] Prognostic models provide a personalised evidence-based approach and may bridge the gap from providing average outcome predictions to individualised outcome predictions. [33] The challenge that clinicians experience in answering patients' prognostic questions, and the availability of an existing large cohort of patients with neck pain in Dutch manual therapists' clinical setting, inspired me to conduct the studies described in this thesis.

Prognostic models can be intended to predict short-term health outcomes or events e.g., immediately after an intervention, or long-term health outcome predictions or events up to several years. A prognostic model may be setting-specific and can be intended for use in neck pain in general or for use in specific subgroups of neck pain with a different prognosis, e.g., people with acute neck pain, persistent neck pain, trauma-related neck pain, work-related neck pain, or patients with radicular pain.

When are prognostic models ready for clinical use?

The evolution of personalised and stratified care coincides with the rapid increase of published multivariable prediction model studies in the various fields of healthcare. Multivariable 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). In this thesis, we focus on prognostic models. Prognostic models are abundant in the medical literature and their appearance increases rapidly in the musculoskeletal literature. [34] Studies of prognostic models comprise three consecutive stages as presented in figure 3: model development (derivation); validation in new settings (external validation); and assessment of a model's clinical impact. [29 32] Model's clinical impact studies are rarely performed, but before a prognostic model can be considered for use in clinical practice, at least successful external validation in a setting comparable to its intended use is needed. [34] Model performance is typically evaluated in terms of discrimination and calibration measures. [29 32]
Discriminative performance indicates whether a model is able to distinguish between patients with and without the outcome of interest and calibration performance refers to the extent of agreement between a model's predicted risks and observed outcomes. [29 32]

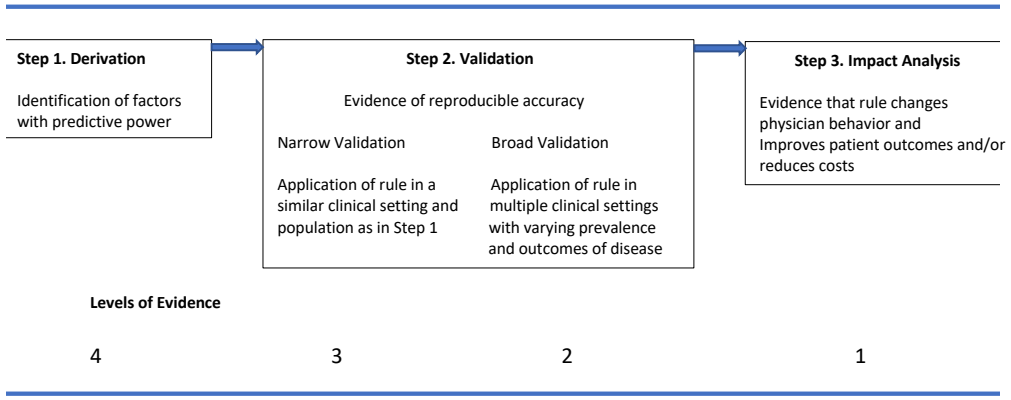


Fig 3. The three consecutive stages of model development [35]

Which outcomes to predict in patients with neck pain?

The construct of recovery in patients with spinal health problems is multidimensional and from a patient perspective may reflect e.g., symptom attenuation, improvement of functional tasks, or achievement of an acceptable quality of life. [36] This is represented in prognostic research studies by various health outcomes such as pain intensity, physical functioning, or health-related quality of life. [37] Outcomes in the musculoskeletal health domain mainly are measured with Patient-Reported Outcome Measures (PROMs) using dichotomized or continuous scores. Outcome measurements that are used in people with neck pain are e.g., the Neck Disability Index (NDI), the Numeric Pain Rating Scale (NPRS), the Visual Analogue Scale (VAS) for pain, the Northwick Park Questionnaire (NPQ), or the Neck Bournemouth Questionnaire (NBQ). [38] [39] There is no clear criterion available for recovery and it is often operationalized in prognostic modelling by dichotomizing PROMs to compare recovered versus non-recovered patients. Different threshold values and parameters -such as the Minimal Important Change (MIC) and the Patient Acceptable Symptom State (PASS) - are used to dichotomize scores or to determine recovery for continuous outcomes.

Aim and outline of this thesis

This thesis aims to improve predictions of recovery of non-specific neck pain in individual patients in primary care with the use of prognostic prediction models by answering two research questions:

(1) Are valid prediction models available for making accurate predictions of recovery in patients with non-specific neck pain?
(2) Can newly developed prognostic models provide accurate predictions of recovery in primary care for patients with non-specific neck pain?

The first research question is addressed in **Chapter 2** with a systematic review of the literature to summarize the risk of bias, applicability, and usability of currently available prognostic prediction models for recovery in patients with non-specific neck pain.

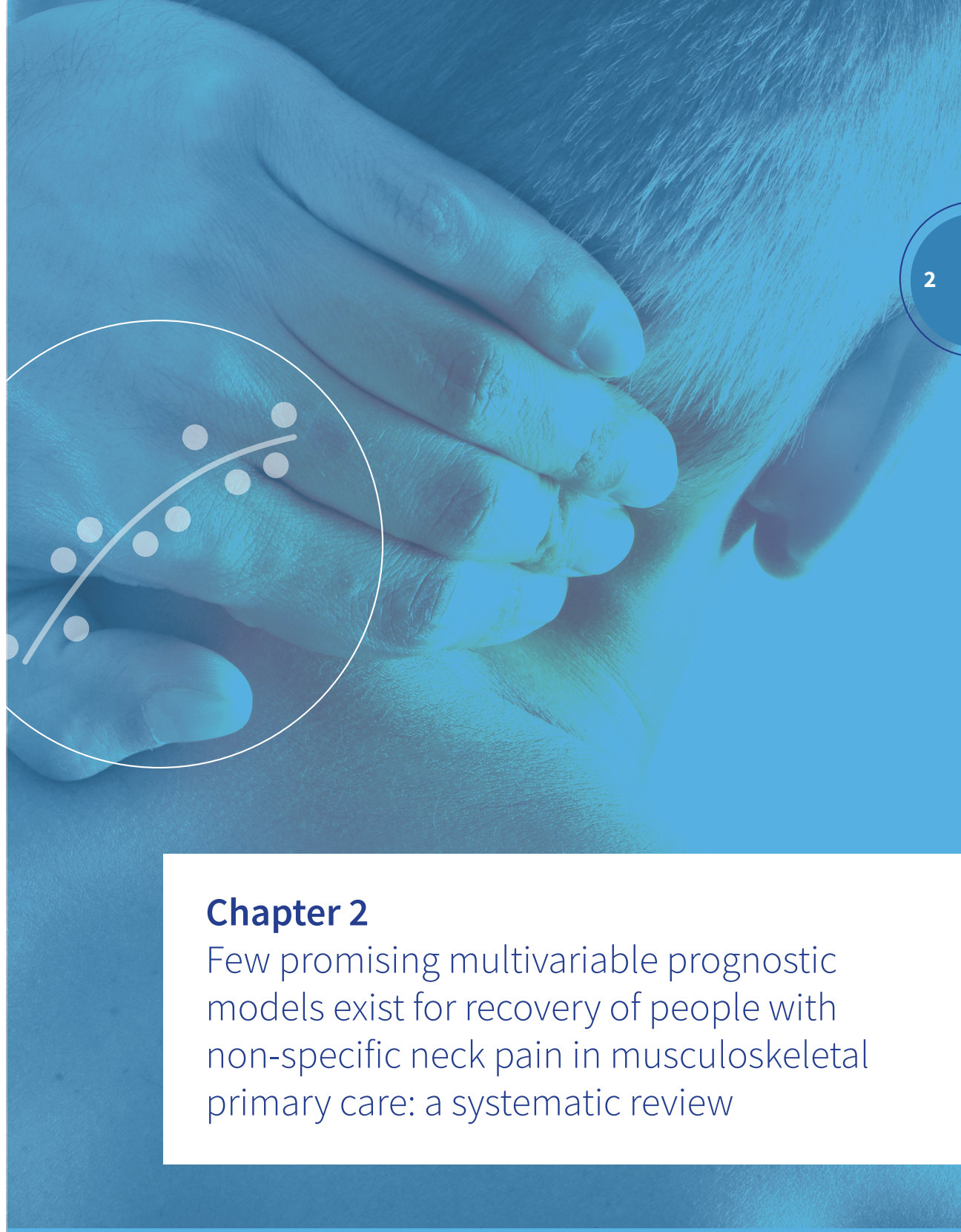
To further address this question, promising models identified in this systematic review are tested in **Chapter 3** for their external validity in an available data set. In **Chapter 4**, common methodological shortcomings in research of prognostic modelling and additional methodological challenges specific to the field of spinal care are discussed and potential solutions are presented.

To address the second research question, in **Chapter 5**, prognostic models that predict 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, are developed and internally validated. The derived post-treatment models are subsequently tested in **Chapter 6** for their external validity and evaluated if they can be updated with new predictors.
In Chapter 7 I summarize the main findings of this thesis and present its limitations. Also, I discuss making prognoses in clinical practice. Additionally, recommendations for future research are provided, question long-term prediction and end with final conclusions.

References

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet. 2020;396(10267):2006-2017. doi:10.1016/S0140-6736(20)32340-0
2. James SL, Abate D, Abate KH, 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. The Lancet. 2018;392(10159): 1789-1858. doi:10.1016/S0140-6736(18)32279-7
3. Safiri S, Kolahi AA, Hoy D, et al. Global, regional, and national burden of neck pain in the general population, 1990-2017: Systematic analysis of the Global Burden of Disease Study 2017. The BMJ. 2020;368. doi:10.1136/bmj.m791
4. James SL, Abate D, Abate KH, 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. The Lancet. 2018;392(10159): 1789-1858. doi:10.1016/S0140-6736(18)32279-7
5. https://www.nivel.nl/nl/nivel-zorgregistraties-eerste-lijn/jaarcijfers-aandoeningen-huisartsenregistraties.
6. van der Dool J, Schermer T. Zorg Door de Fysiotherapeut; Jaarcijfers 2018 En Trendcijfers 2014 – 2018.; 2019.
7. Veldkamp R, Kruisselbrink M, Meijer W, Zorgregistraties N, Lijn E. Zorg Door de Fysiotherapeut. Nivel Zorgregistraties Eerste Lijn, Jaarcijfers 2020 En Trendcijfers 2017-2020.; 2022. https://www.nivel.nl/nl/nivel-zorgregistraties-eerste-lijn/nivel-zorgregistraties-eerste-lijn.
8. Hoy D, March L, Woolf A, et al. The global burden of neck pain: Estimates from the global burden of disease 2010 study. Annals of the Rheumatic Diseases. 2014;73(7):1309-1315.
9. Gebhart GF, Schmidt RF, eds. Encyclopedia of Pain. Springer Berlin Heidelberg; 2013. doi:10.1007/978-3-642-28753-4
10. Vos CJ, Verhagen AP, Passchier J, Koes BW. Clinical course and prognostic factors in acute neck pain: an inception cohort study in general practice. Pain Medicine. 2008;9(5):572-580 9p. doi:10.1111/j.1526-4637.2008.00456.x
11. Parikh P, Santaguida P, Macdermid J, Gross A, Eshtiaghi A. Comparison of CPG's for the diagnosis, prognosis and management of non-specific neck pain: a systematic review. BMC Musculoskeletal Disorders. 2019;20(1):81. doi:10.1186/s12891-019-2441-3
12. Borghouts JAJ, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. Pain. 1998;77(1):1-13.
13. Haldeman S, Carroll L, Cassidy JD, Schubert J, Nygren Å. The Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. European Spine Journal. 2008;17(S1):5-7. doi:10.1007/s00586-008-0619-8
14. Verhagen AP. Physiotherapy management of neck pain. Journal of Physiotherapy. Published online 2021. doi:10.1016/j.jphys.2020.12.005
15. Sterling M, de Zoete RMJ, Coppieters I, Farrell SF. Best Evidence Rehabilitation for Chronic Pain Part 4: Neck Pain. Journal of Clinical Medicine. 2019;8(8):1219. doi:10.3390/jcm8081219
16. Verhagen AP, Lewis M, Schellingerhout JM, et al. Do whiplash patients differ from other patients with non-specific neck pain regarding pain, function or prognosis? Manual Therapy. 2011;16(5):456-462. doi:10.1016/j.math.2011.02.009
17. Stenneberg MS, Rood M, Bie R de, Schmitt MA, Cattrysse E, Scholten-peeters GG. To What Degree Does Active Cervical Range of Motion Differ Between Patients With Neck Pain , Patients With Whiplash, and Those Without Neck Pain ? A Systematic Review and Meta-Analysis. Archives of Physical Medicine and Rehabilitation. Published online 2016. doi:10.1016/j.apmr.2016.10.003
18. Hush JM, Lin CC, Michaleff Z a, Verhagen A, Refshauge KM. Prognosis of Acute Idiopathic Neck Pain is Poor: A Systematic Review and Meta-Analysis. Archives of Physical Medicine and Rehabilitation. 2011;92(5):824-829. doi:10.1016/j.apmr.2010.12.025
19. 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(8):1237-1244. doi:10.1016/j.pain.2013.03.032
20. Carroll LJ, Hogg-Johnson S, van der Velde G, et al. Course and Prognostic Factors for Neck Pain in the General Population. Spine (Phila Pa 1976). 2008;33(Supplement):S75-S82. doi:10.1097/BRS.0b013e31816445be
21. Carroll LJ, Holm LW, Hogg-Johnson S, et al. Course and prognostic factors for neck pain in whiplash-associated disorders (WAD): results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. J Manipulative Physiol Ther. 2009;32(2 Suppl):S97--S107. doi:10.1016/j.jmpt.2008.11.014
22. Nordin M, Carragee EJ, Hogg-Johnson S, et al. Assessment of neck pain and its associated disorders: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. J Manipulative Physiol Ther. 2009;32(2 Suppl):S117-40. doi:10.1016/j.jmpt.2008.11.016
23. Corp N, Mansell G, Stynes S, et al. Evidence-based treatment recommendations for neck and low back pain across Europe: A systematic review of guidelines. European Journal of Pain (United Kingdom). 2020;(March):1-21. doi:10.1002/ejp.1679
24. Guerrero AVS, Setchell J, Maujean A, Sterling M. A Comparison of Perceptions of Reassurance in Patients with Nontraumatic Neck Pain and Whiplash-Associated Disorders in Consultations with Primary Care Practitioners—An Online Survey. Pain Medicine (United States). 2020;21(12):3377-3386. doi:10.1093/PM/PNAA277
25. Gordon Guyatt, Drummond Rennie, Maureen O. Meade DJC. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. 3rd ed. McGraw-Hill Education; 2015.
26. Cook CE, Moore TJ, Learman K, Showalter C, Snodgrass SJ. Can experienced physiotherapists identify which patients are likely to succeed with physical therapy treatment? Archives of Physiotherapy. 2015;5(1):3. doi:10.1186/s40945-015-0003-z
27. Newell D, Field J, Visnes N. Prognostic accuracy of clinicians for back, neck and shoulder patients in routine practice. Chiropractic and Manual Therapies. 2013;21(1). http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L52896683
28. Dagfinrud H, Storheim K, Magnussen LH, et al. The predictive validity of the Örebro Musculoskeletal Pain Questionnaire and the clinicians' prognostic assessment following manual therapy treatment of patients with LBP and neck pain. Manual Therapy. 2013;18(2):124-129. doi:10.1016/j.math.2012.08.002
29. Riley RD, Van Der Windt DA, Croft P, Moons KGM. Prognosis Research in Healthcare. first. Oxford University Press; 2019.
30. Moons KGM, Altman DG, Reitsma JB, Collins GS. New Guideline for the Reporting of Studies Developing, Validating, or Updating a Prediction Model. Clinical Chemistry. 2015;61(3):565-566. doi:10.1373/clinchem.2014.237883

31. Mansell G, Corp N, Wynne-Jones G, Hill J, Stynes S, Windt D. Self-reported prognostic factors in adults reporting neck or low back pain: An umbrella review. European Journal of Pain. 2021;25(8):1627-1643. doi:10.1002/ejp.1782
32. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer Science and Business Media,; 2019.
33. Riley RD, van der Windt DA, Croft P, Moons KGM. Prognosis Research in Health Care, Concepts, Methods, and Impact. 1st ed. Oxford University Press; 2019.
34. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Medicine. 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381
35. Mcginn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users ' Guides to the Medical Literature XXII : How to Use Articles About Clinical Decision Rules. JAMA : the journal of the American Medical Association. 2015;284(1):79-84.
36. 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 & Research. 2008;61(1):124-131. doi:10.1002/art.24162
37. Chiarotto A, Deyo RA, Terwee CB, et al. Core outcome domains for clinical trials in non-specific low back pain. European Spine Journal. 2015;24(6):1127-1142. doi:10.1007/s00586-015-3892-3
38. Bobos P, MacDermid JC, Walton DM, Gross A, Santaguida PL. Patient-Reported Outcome Measures Used for Neck Disorders: An Overview of Systematic Reviews. Journal of Orthopaedic & Sports Physical Therapy. 2018;48(10):775-788. doi:10.2519/jospt.2018.8131
39. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: A systematic review. Quality of Life Research. 2012;21(4):659-670. doi:10.1007/s11136-011-9965-9



2

Chapter 2

Few promising multivariable prognostic models exist for recovery of people with non-specific neck pain in musculoskeletal primary care: a systematic review

Chapter 2. Few promising multivariable prognostic models exist for recovery of people with non-specific neck pain in musculoskeletal primary care: a systematic review

Roel W. Wingbermühle, Emiel van Trijffel, Paul M. Nelissen, Bart Koes, Arianne P. Verhagen

Journal of Physiotherapy. 2018 Jan; 64 (1): 16-23

Question: Which multivariable prognostic model(s) for recovery in people with neck pain can be used in primary care? **Design:** Systematic review of studies evaluating multivariable prognostic models. **Participants:** People with non-specific neck pain presenting at primary care. **Determinants:** Baseline characteristics of the participants. **Outcome measures:** Recovery is measured as pain reduction, reduced disability, or perceived recovery at short-term and long-term follow-up. **Results:** Fifty-three publications were included, of which 46 were derivation studies, four were validation studies and three concerned combined studies. The derivation studies presented 99 multivariate models, all of which were at high risk of bias. Three externally validated models generated usable models in low risk of bias studies. One predicted recovery in non-specific neck pain, while two concerned participants with whiplash-associated disorders (WAD). Discriminative ability of the non-specific neck pain model was the area under the curve (AUC) 0.65 (95% CI 0.59 to 0.71). For the first WAD model, discriminative ability was AUC 0.85 (95% CI 0.79 to 0.91). For the second WAD model, specificity was 99% (95% CI 93 to 100) and sensitivity was 44% (95% CI 23 to 65) for prediction of non-recovery, and 86% (95% CI 73 to 94) and 55% (95% CI 41 to 69) for prediction of recovery, respectively. Initial Neck Disability Index scores and age were identified as consistent prognostic factors in these three models. **Conclusion:** Three externally validated models were found to be usable and to have a low risk of bias, of which two showed acceptable discriminative properties for predicting recovery in people with neck pain. These three models need further validation and evaluation of their clinical impact before their broad clinical use can be advocated. **Registration:** PROSPERO CRD42016042204.
[Wingbermühle RW, van Trijffel E, Nelissen PM, Koes B, Verhagen AP (2018) Few promising multivariable prognostic models exist for recovery of people with nonspecific neck pain in musculoskeletal primary care: a systematic review. Journal of Physiotherapy 64:16–23]

Introduction

Globally, neck pain is one of the main contributors to years lived with disability.[1,2] Improvements in pain and disability typically occur in the first weeks after the onset of an episode of neck pain, but residual pain and disability beyond this time are often of substantial severity and persist for at least 1 year.[3] High baseline neck pain intensity and disability scores have been identified as predictors for poor outcome in people with neck pain.[4] Cost-effectiveness and short-term beneficial effects of non-invasive primary care treatment have been reported, while long-term effects are still limited.[5–8] Subgrouping of people with neck pain based on their prognosis may enhance treatment outcomes by

enabling tailored treatment and management strategies.[9–11] High-quality research on neck pain prognosis has been a research priority for over a decade.[12]
A fundamental shift in clinical practice has been proposed towards the prospective relationships between phenotypic, genomic, and environmental assessment of patients.[13] It is argued that prognostic profiles allow a more holistic view and can better manage subjectively reported health problems than diagnostic labels.[13] These prognostic profiles should also more accurately mirror daily practice.[14]

Prognostic factors can be developed based on demographic factors, disease characteristics, or factors derived from history taking, physical examination, or additional examinations (such as imaging, blood assays, urine tests, or other biological measurements).[15] Multiple factors are likely to interact with each other, so multivariable prognostic models that consider correlations between predictors have been proposed.[4,16–18] Development of multivariable prognostic models consists of three consecutive stages: developing the model (derivation); validating its performance in new patients (external validation); studying its clinical impact (impact analysis).[17,19]
Numerous multivariable prognostic models in musculoskeletal primary care for people with neck pain have been developed. To our knowledge, these models have not been evaluated systematically using tools specifically designed to assess the quality and usability of primary multivariable prognostic model studies included in a systematic review.
Several systematic reviews have been conducted to summarise the value of prognostic models in the musculoskeletal domain, [20–22] with one focusing on neck pain alone.[23] These reviews concluded that the methodological quality of the included studies was often poor to moderate, validation studies are rare, and routine clinical use is therefore not supported. Methodological quality was assessed in these systematic reviews using tools not specifically designed for assessing the quality of prediction models. Only recently, PROBAST (Prediction model study Risk Of Bias Assessment Tool) has become available; it is designed to assess the risk of bias and concerns about applicability of studies that develop and/or validate a multivariable prediction model when they are included in systematic reviews.[24–26]
To our knowledge, no systematic review on multivariable prognostic models for recovery (pain reduction, reduced disability, or perceived recovery) of people of all ages presenting in primary care with neck pain has been conducted using an up-to-date methodology.
This systematic review aimed to summarise the validity and applicability of multivariable prognostic models for recovery in people with neck pain in primary care. Therefore, the specific research question for this systematic review was:

Which multivariable prognostic model(s) for recovery in people with neck pain can be used in primary care?

Method

Identification and selection of studies
MEDLINE, EMBASE, and CINAHL databases were searched to retrieve all relevant studies on multivariable prognostic models for recovery of neck pain from inception up to May 3,

2016. This search was based on a validated strategy adapted for this study.[20,27,28] The full search strategy is listed in Appendix 1. De-duplication was performed in Mendeley and hand checked.[29] No language restrictions were imposed. Additional manual searching of reference lists of all included studies was performed. To be eligible for inclusion, studies had to generate multivariable prognostic models using data from prospective cohort studies and randomised, controlled trials on participants of any age with non-serious specific and non-specific neck pain. Models in all stages of their development were considered. Models were defined as those constructed by multivariable analysis from a combination of at least two predictors associated with a particular outcome, while derived models could contain one remaining variable.[17,30,31,32] All baseline characteristics that are feasible to measure in primary care were considered as potential predictors. Studies were included when the outcome concerned pain reduction, reduced disability, or perceived recovery at any time of follow-up. The inclusion criteria are summarised in Box 1. Studies aimed at (cost-) effectiveness, side effects, or developing a questionnaire were excluded. Studies using clinical procedures involving skin penetration like injection, acupuncture, or dry needling were also excluded.

Two reviewers (RW, PN) independently screened records for possibly relevant studies based on title and abstract. Subsequently, full texts of potentially relevant articles were independently assessed for eligibility. Discrepancies between reviewers were resolved through discussion or by a third reviewer (APV).

Box 1. Inclusion criteria.

Models
- Constructed with multivariable analysis
- Combination of at least two predictors
- Any stage of development

Design
- Prospective cohort studies
- Randomised, controlled trials

Participants
- People of any age
- Non-serious specific or non-specific neck pain at any stage[a]

Determinants
- Baseline characteristics at intake
- Applicable to and easily obtained in non-invasive musculoskeletal primary care

Outcome to be predicted
- Pain
- Disability
- Perceived recovery

[a] Neck pain was defined as pain located in the anatomic region of the neck from the linea nuchea superior to the spina scapula, with or without radiation to the trunk or upper limb.[33,34] Non-specific neck pain was defined as neck pain without an identified pathological basis. Non-serious neck pain was defined as neck pain with an identified pathological basis, but with no contra-indication for musculoskeletal primary care.[35]

Assessment of characteristics of studies

Quality

The Quality of the selected studies was assessed using the pre-publication version of PROBAST.[36] PROBAST was developed using a Delphi process involving 40 experts in the fields of systematic review methodology and prediction research. It was designed to assess the risk of bias, applicability, and usability of multivariable prediction model studies included in a systematic review using a similar domain-based approach as the revised tool for the quality assessment of diagnostic accuracy studies (QUADAS-2). Judgements on high, low, or unclear risk of bias for reported estimates of the model's predictive performance were made for five key domains (participant selection, predictors, outcome, sample size and participant flow, and analysis) after judgement of signalling questions. As the signalling question was to determine whether there was a reasonable number of outcome events in logistic regression, the number of events in the smallest group was divided by the total degrees of freedom used during the whole modelling process. Counting degrees of freedom was based on each time a variable or its category was tested on the outcome. Univariable predictors were considered here as part of the whole modelling process if they were selected based on their p-value. Rating was according to the 'rule of thumb' of 10 events per variable.[37] For linear regression, the number of participants was divided by the number of predictors. High, low, or unclear concerns about applicability regarding the review question were made in a similar structure for three key domains (participant selection, predictors, and outcome).
An overall judgement about the risk of bias and applicability of the prediction model evaluation was reached based on a separate summative rating across all domains for derivation and validation studies according to the PROBAST criteria. Finally, a model's usability was rated for its presentation with sufficient detail to be used in the intended context and target population.

Two reviewers (RW, PN) independently assessed the quality of the selected studies. Discrepancies and unclear items were resolved through discussion or, if necessary, adjudication by a third reviewer (APV). Percentage agreement and Cohen's kappa in a 2x2 contingency table were used to describe the level of agreement between the two reviewers for the judgements of the risk of bias and applicability domains. For this purpose, 'high' and 'unclear' ratings were collapsed into one category. Rating of models within the same study were combined into one variable per reviewer if ratings were the same.

Data extraction

In accordance with the CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) checklist for prediction model development studies, data were extracted from the included studies on study setting, country, and dates; participants' condition and treatment; number of variables and events; predictors in the model; predicted outcome and follow-up; model performance and stage; clinical measures; and model presentation.[38] Data extraction was performed independently by two reviewers (RW, PN), and randomly cross-checked by a third reviewer (APV).

Data analysis and evaluation

A qualitative synthesis was performed to evaluate whether a model was ready for clinical use by analysing the model's risk of bias, applicability, and usability as related to its performance accuracy. Analyses were conducted separately for derivation studies and validation studies. For subdividing the studies according to study stage, validation performed with non-random split data (type 2b) was considered to be external validation.[18,39] A model was judged to be ready for clinical use if it was usable and externally validated in a study with an overall low risk of bias, while showing acceptable discriminative performance. Prediction models were accurate if they were able to discriminate between people with and without the outcome.[40] Model discriminative performance was considered acceptable if the area under the curve (AUC) of the receiver operating characteristic (ROC) curve for continuous outcome or c-statistic for binary outcome was ≥ 0.7.[41] The ROC curve plots the model's true-positive prediction rate (sensitivity) versus the false-positive prediction rate (one minus the specificity) over all possible discrimination thresholds of predicted probability of the occurrence of the outcome. The c-statistic is comparable to the AUC for binary outcome and is the proportion of pairs – one individual with and one individual without the outcome – in which the individual who experienced the outcome had a higher probability of experiencing the outcome than the individual who did not experience the outcome, as predicted by the model.[40] In addition, we searched for prognostic factors consistently appearing in final models from low risk of bias studies.

Results

The flow of studies through the review

Searching MEDLINE, EMBASE, and CINAHL initially yielded 1119, 1554, and 143 records, respectively. After the removal of duplicate citations, 2398 remained. Of these, 2305 records were excluded based on title and abstract. Hand searching added five potentially relevant publications, so a total of 98 full-text articles were evaluated for eligibility. Forty-five studies, of which 27 did not involve multivariable analysis, were excluded. Fifty-three studies met the selection criteria; 46 of these were derivation studies, while four were validation studies only, and three combined derivation and validation in one publication (Figure 1).

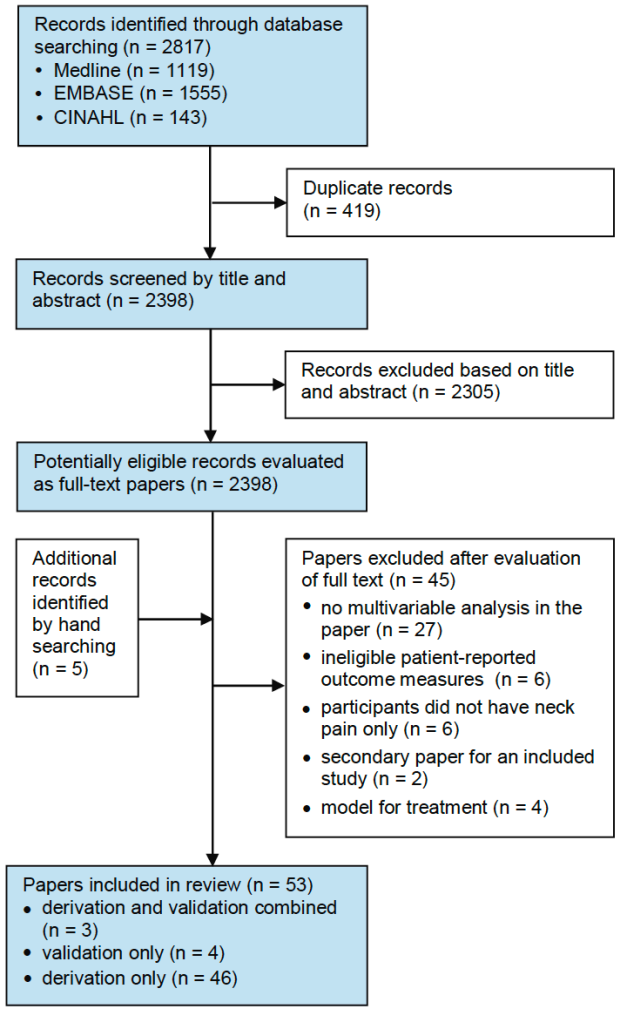


Figure 1. Flow of studies through the review.

Characteristics of included studies

The characteristics of the 46 included derivation studies are presented in Appendix 2. The characteristics of the four validation studies and the three combined studies are presented in Appendix 3.

Derivation studies

The mean age of participants in the derivation studies ranged from 30 to 65 years. Mean symptom duration at baseline ranged from 60 days to 108 months. Follow-up for outcome measurement among the included derivation studies ranged between 1 week and 5 years. Outcomes were measured using various patient-rated disability scales, global rating of change, or pain scales. In total, 99 models were derived in 49 studies (excluding two models newly developed in a validation study).[42]

Twenty-six studies described 58 models concerning participants with Whiplash-Associated Disorders (WAD); 35 were acute, three were subacute, six chronic, and 14 were of any or of unknown duration. Twenty-three studies described 41 models concerning participants with neck pain conditions; three acute, five subacute, six chronic, four with or without arm symptoms, two nerve-related arm pain, and 21 of any duration or unknown. The number of predictors in the final models varied from 1 to 10. The included derivation studies assessed a variety of types of predictors, such as history variables (eg, age, gender, pain/symptoms, symptom duration, disability, psychosocial, contextual) and physical examination variables (eg, range of motion, pain provocation, pain or temperature threshold). Twelve models were presented as a score chart, nomogram, prediction or decision rule.

Quality

Judgements about the risk of bias, applicability, and usability are shown in Tables 1 and 2. Agreement between the two reviewers for judgements of the five risk of bias domains (participant selection, predictors, outcome, sample size and participant flow, and analysis) was 71, 69, 51, 98 and 92%, respectively. In the outcome domain, reviewers disagreed mainly due to their interpretation of the impact of predictors that were not excluded from the outcome definition. Agreement between the two reviewers for judgements of the three applicability domains (participant selection, predictors, and outcome) was 74, 90 and 84%, respectively. In two instances, the third reviewer had to make a decision. Cohen's kappa appeared not applicable, due to consistent very low or zero prevalence. All 49 studies had a high risk of bias and every study had a high risk of bias in the sample size and participant flow domain, while 43 were biased in the analysis domain. In 42 studies, models were judged to have low concerns regarding their applicability. Four studies contained 11 models with a reasonable number of outcome events according to the definition based on events per variable or participants per predictor.[43–46] All enrolled participants were included in the analysis of nine studies for 12 models.[47–55] Missing data were handled appropriately in seven studies for 17 models.[51,53,56–60] Two derivation studies performed internal validation.[61,62]

The model's overall performance was described in 34 studies by some form of R-squared statistic (R^2). In 11 studies, calibration and/or discrimination measures were described for 19 models. Two studies checked internal validity by cross-validation bootstrapping; one of them computed a shrinkage factor.[61,62] Some form of treatment was performed in 29 studies, of which eight described that participants received a specific therapy, like manual therapy, a multi-modal program, standardised physiotherapy, or neural tissue management.

Table 1
Risk of bias, applicability, and model usability among the studies with development models.[11,43–71,75,82–99]

Study	Risk of bias, signalling questions					Applicability			Overall judgement		
	Participant selection	Predictors	Outcome	Sample size and participant flow	Analysis	Participant selection	Predictors	Outcome	Risk of bias	Applicability	Usability
Angst[56]	L	L	L	H	H	L	H	L	H	H	Y
Åsenlöf[68]	L	H	L	H	L	L	L	L	H	L	N
Atherton[82]	L	L	L	H	H	L	L	L	H	L	N
Baltov[69]	U	L	U	H	H	L	L	L	H	L	N
Bohman[61]	L	L	L	H	U	L	L	L	H	L	N
Buitenhuis[83]	L	L	U	H	H	L	L	L	H	L	Y
Bunketorp[84]	L	H	L	H	H	L	L	L	H	L	N
Cai[47]	L	U	H	H	H	L	L	L	H	L	Y
Carstensen[64]	L	L	L	H	L	L	L	L	H	L	N
Cecchi[85]	L	L	1U/1L	H	H	L	L	L	H	L	N
Chiarotto[57]	L	L	U	H	H	L	L	L	H	L	Y
Cleland[11]	L	L	L	H	H	L	L	L	H	L	Y
Cleland[86]	L	L	L	H	H	L	L	L	H	L	Y
Cobo[43]	L	L	L	H	H	L	L	L	H	L	Y
Dagfinrud[70]	L	U	L	H	H	L	L	L	H	L	N
Gun[87]	L	L	L	H	H	L	L	L	H	L	N
Hanney[48]	L	L	L	H	H	L	L	L	H	L	Y
Hartling[88]	L	L	L	H	H	L	L	L	H	L	Y
Hendriks[63]	L	L	L	H	H	L	L	L	H	L	N
Hill[49]	L	L	L	H	H	L	L	L	H	L	N
Hoving[58]	L	L	L	H	H	L	L	L	H	L	N
Keating[75, a]	U	L	L	H	H	U	L	L	H	U	N
Kjelmann[89]	L	L	L	H	H	L	L	L	H	L	N
Kyhlbäck[44]	L	U	L	H	H	L	L	L	H	L	N
Landers[50]	L	U	L	H	H	L	L	L	H	L	Y
Lankester[45]	U	U	U	H	H	L	L	U	H	U	N
Michaelson[90]	L	L	L	H	H	L	L	L	H	L	N
Nederhand[65]	L	L	L	H	H	L	L	L	H	L	Y
Nee[66]	L	L	L	H	U	L	L	L	H	L	Y
Nieto[91]	L	L	U	H	H	L	L	L	H	L	N
Pape[92]	H	L	L	H	H	U	L	L	H	L	N
Peterson[67]	U	L	H	H	H	L	2U/2L	L	H	H/L	N
Pool[93]	L	L	L	H	H	L	L	L	H	L	N
Puentedura[51]	L	L	L	H	H	L	L	L	H	L	Y
Radanov[52, b]	U	U	H	H	H	U	U	U	H	U	Y
Raney[94]	L	L	L	H	H	L	L	L	H	L	Y
Rebbeck[95]	L	L	U	H	H	L	L	L	H	L	N
Ritchie[96]	L	L	L	H	H	L	L	L	H	L	Y
Rubinstein[59]	L	L	L	H	H	L	L	L	H	L	N
Schellingerhout[53, b]	L	L	L	H	L	L	L	L	H	L	Y
Saavedra-Hernández[54]	L	L	L	H	H	L	L	L	H	L	Y
Sterling[62]	L	L	L	H	U	L	L	L	H	U	1Y/2N
Sterling[71]	L	L	L	H	H	L	L	L	H	U	1Y/2N
Sterner[46]	L	U	U	U	H	L	L	L	H	L	N
Sturzenegger[97]	L	L	U	H	H	L	L	L	H	L	N
Tseng[55]	L	L	L	H	H	L	L	L	H	L	Y
Vos[60]	L	L	L	H	H	L	L	L	H	L	N
Walton[98]	L	U	L	H	H	L	L	L	H	L	Y
Williamson[99]	L	L	L	H	H	L	L	L	H	L	Y

H = high, L = low, N = no, U = unclear, Y = yes.
Green shading = favourable result, Yellow shading = unclear or mixed results, Red shading = unfavourable result.
[a] Type 2b study, intermediate (temporal) validation.
[b] Type 3 study, development and validation using separate data set.

Table 2
Risk of bias, applicability, and model usability among the studies with validation models.[42,52,53,72–75]

Study	Risk of bias, signalling questions					Applicability			Overall judgement		
	Participant selection	Predictors	Outcome	Sample size and participant flow	Analysis	Participant selection	Predictors	Outcome	Risk of bias	Applicability	Usability
Cleland[72, c]	L	L	L	U	U	L	L	L	H	L	Y
Fritz[73, c]	L	L	L	U	H	L	L	L	H	L	Y
Keating[75, a]	U	L	L	U	H	U	L	L	H	U	N
Radanov[52, b]	U	U	H	H	H	U	U	U	H	U	Y
Ritchie[74, c]	L	L	L	L	L	L	L	L	L	L	Y
Schellingerhout[53, b]	L	L	L	L	L	L	L	L	L	L	Y
Sterling[42, c/d]	L	L	L	L	L	L	L	L	L	U	Y

H = high, L = low, N = no, U = unclear, Y = yes.
Green shading = favourable result, Yellow shading = unclear or mixed results, Red shading = unfavourable result.
[a] Type 2b study, intermediate (temporal) validation.
[b] Type 3 study, development and validation using separate data set.
[c] Type 4 study, validation only.
[d] Contains two regression models developed in validation study.

Performance

Seven models reported discriminative ability (AUC or c-statistic) ranging from 0.66 to 0.93.[53,61,63–67] The number of events per variable was > 5 in two of these studies,[53,64] one of which was subsequently validated and upheld its model performance.[53] Ten studies presented 15 models with an R^2 or adjusted $R^2 \geq 0.5$.[11,50,51,56,62,63,68–71] For two of these models, external validation studies were subsequently performed,[42,72] one of which concluded that the model could not be validated.[72]

Validation studies

Among the validation studies, the sample size ranged from 16 to 315 and the mean age of participants ranged from 32 to 49 years. Outcomes were measured between 1 week and 12 months, mostly with the Neck Disability Index (NDI) scale or Global Rating of Change.
One study concerned an insurance company population[52] and six studies concerned populations from physiotherapy care, four of which combined a physiotherapy setting with other settings.[42,53,73,74]
In two studies, models were tested in a different country than the derivation study.[42,53]
Four studies contained models on neck pain,[53,72,73,75] while three studies concerned models for WAD.[42,52,74] Two studies reported that the models could not be validated,[72,73] and one study reported no improvement based on positive predictive value and only weak improvement based on negative predictive value.[75] Two studies reported support for their models based on model performance measures.[53,74] One study reported support based on percentage correct predictions only and did not give any model performance measures.[52]
One WAD study concluded that the model was not accurate because it overestimated the NDI score, and reported discriminative ability if the outcome was dichotomised.[42]

Quality

Agreement between the two reviewers for judgements of the five risk of bias domains (participant selection, predictors, outcome, sample size and participant flow, and analysis) was 57, 86, 57, 71 and 85%, respectively. In the participant selection and outcome domains, reviewers disagreed mainly due to their interpretation of the impact of selection criteria and predictors that were not excluded from the outcome definition. Agreement between the two reviewers for the three applicability domains (participant selection, predictors, and outcome) was 86, 71 and 100%, respectively. Cohen's kappa appeared not to be applicable, due to consistent very low or zero prevalence. Four studies had an overall high risk of bias in one or more domains;[52,72,73,75] among these studies, two models were judged as having unclear concerns regarding applicability[52,75] and one was judged as not usable.[75]
One study performed type 2b non-random split validation.[75] High risk of bias was consistent in the analysis domain, mostly due to dichotomised variables and lack of information. Three studies with a low risk of bias generated usable models.[42,53,74] Two of these models were judged to have low concerns regarding their applicability.[53,74]

Performance

In the three validation studies with a low risk of bias overall, one model was intended for use in people with non-specific neck pain,[53] while two concerned people with WAD.[42,74]
Discriminative ability of the non-specific neck pain model was AUC 0.65 (95% CI 0.59 to 0.71) and that of the corresponding score chart was 0.66 (95% CI 0.59 to 0.72).[53] Applicability concerns were low and the score chart was clinically usable. Discriminative ability of the first WAD model was AUC 0.85 (95% CI 0.79 to 0.91), and for calibration, the study reported an overestimation of the NDI outcome.[42] This study did not recalibrate the validated model but used its predictors for developing a new model, presenting AUC 0.89 (95% CI 0.84 to 0.94) and 0.91 (95% CI 0.86 to 0.95) if adjusted for study site. The second WAD study tested a prediction rule for two of its three recovery pathways, one moderate to severe path with an outcome of NDI ≥ 30%, and one full recovery path with an outcome of NDI ≤ 10%.74 For the path of NDI ≥ 30%, specificity was 99% (95% CI 93 to 100) and sensitivity was 44% (95% CI 23 to 65). For the path of NDI ≤ 10%, specificity was 86% (95% CI 73 to 94) and sensitivity was 55% (95% CI 41 to 69). Applicability concerns were low and the model was clinically usable.
Consistent prognostic factors in these three models were age, and initial NDI score for WAD. Age lost its significance initially during a low risk of bias derivation study but it regained significance after adjusting for research site.[42]

Discussion

This systematic review included 53 studies of 99 derivation models and seven models tested for validation for the prediction of recovery in people with neck pain. Two WAD models and one nonspecific neck pain model were found to be promising for use in primary care settings. These findings are in line with previous systematic reviews on prognostic models for neck pain recovery. One review included six studies and concluded that most models were in the developmental stage, often with moderate study quality.[20] Another review on clinical prediction rules included 18 studies with four models at the derivation stage and

no neck pain models appearing validated.[21] A second review of clinical prediction rules concluded that two out of the three neck pain studies met their quality criteria. However, quality criteria for prognostic studies were used instead of ones specifically developed for prognostic models.[22] The most recent review on clinical prediction rules for prognosis and treatment prescription in neck pain found that 11 out of 15 clinical prediction rules were at the initial stage of development and seven models had undergone validation.[23] All previous reviews concluded that the methodological quality of the original studies was generally low and few models had undergone validation. Therefore, broad routine clinical use was not recommended yet, which was a conclusion shared with other reviews within the spinal musculoskeletal field.[20,76,77]

Evaluating the studies with up-to-date criteria using the PROBAST tool, a large number of derivation studies with high risk of bias was found, especially in the analysis and sample size/participant flow domains. Studies with a high risk of bias may find inflated discriminative performance. Reporting and methodological standards were often not met, for instance, with respect to reporting of missing data and model performance measures (eg, calibration, discrimination), appropriate handling of missing data (eg, multiple imputation), or correction for overfitting (eg, bootstrapping, shrinkage). Overfitting is one of the biggest concerns and occurs when too many predictors are included in the analysis, especially in small data sets resulting in derived models fitting the data too closely.[78] In that case, the model could obtain idiosyncratic features that are specific to the derivation data itself, resulting in a model that predicts accurately in a derivation sample but performs poorly when applied to other individuals.[38] Too many predictors and categorical variables were often selected in derivation studies and the sample size became very low, resulting in high risk for overfitting. Few studies corrected for overfitting using techniques such as bootstrapping and shrinkage. To reduce overfitting, it is recommended that future researchers collect more data, if possible, select predictors based on former knowledge, and use bootstrapping and shrinkage techniques.[78]

This is the first study that systematically evaluated multivariable prognostic models for recovery of people of any age presenting in primary care with neck pain, using a tool specifically designed for assessing the risk of bias and applicability of prognostic model studies. Using PROBAST – instead of tools not specifically designed for assessing prognostic model studies – facilitates evaluating items specific to prognostic models such as overfitting, data complexities, and a model's performance. However, PROBAST does not provide a guideline for scoring items as yet and we had to construct our own. For example, we interpreted the signalling question on the reasonable number of outcome events on the 'rule of thumb' of 10 events per variable; this was rigorous because it was based on the degrees of freedom used. A less rigorous interpretation would probably result in the review spuriously concluding lower risk of bias for the derivation studies.

Another limitation was that WAD studies were included with populations that included primary care patients, people recruited from hospital emergency departments and recruited via general advertisements. It might be possible that predictors for recovery differ between patients in primary care versus emergency departments, or the general population. Another potential limitation could have been publication bias. Although a large number of studies without language restriction were included, no non-English studies were obtained, which may have potentially yielded more negative results.

The vast majority of the models cannot be used in a clinical situation yet, because their derivation studies had a high risk of bias and validation was not executed or unsuccessful. Nevertheless, this review found three validated models that are considered to be promising and may provide support for clinicians in their decision-making process.

The Ritchie two-way WAD model predicted full recovery by NDI ≤ 32% and age ≤ 25 years, and ongoing moderate/severe disability by NDI ≥ 40%, age ≥ 35 years, and hyperarousal (Posttraumatic Diagnostic Scale subscale ≥ 6).[74] The Sterling WAD model predicted disability by initial NDI, age, left rotation range of motion, cold pain threshold, Impact of Events Scale, and blood flow (Quotient of Integrals).[42] The Schellingerhout non-specific neck pain model predicted recovery by age, pain intensity, headache, radiation to elbow/shoulder, previous neck complaints, low back pain, employment status, and quality of life (EuroQOL).[53]

Baseline disability appeared to be a consistent prognostic factor in WAD and could support treatment decision-making because disability can effectively be reduced by primary care interventions in WAD and neck pain.[79,80]

Rather than the development of new models, (further) validating, adjusting, or updating existing (high-quality) models is advocated.[19,81] For the three promising models, further validation and evaluation of clinical impact is advised before their broad clinical use can be advocated. The neck pain model showed a small pre-test to post-test probability shift, and testing the model or its chart in a comparable setting with other prevalence rates is recommended. Further, testing the performance of the two WAD models in a primary care setting alone is required.

What is already known on this topic: Improvements in pain and disability typically occur in the first weeks after the onset of an episode of neck pain, but residual pain and disability beyond this time are often of substantial severity and persist for at least 1 year. Subgrouping people with neck pain based on their prognosis may enhance treatment outcomes by enabling tailored treatment and management strategies.

What this study adds: Although many models have been developed and investigated for their ability to predict recovery of people with neck pain, few are suitable to use. However, two models for whiplash-associated disorders and one model for non-specific neck pain were found to be suitable for use in primary care settings.

Appendix 1. Search strategy

Pubmed:

(I) #1 multivariable prognostic models van Oort:

(Decision Support Techniques[Mesh] OR Predictive Value of Tests[Mesh] OR clinical prediction[tiab] OR prognos*[tiab] OR predict*[tiab])

(I) #4 primary musculoskeletal care FT van Oort:

(Primary Health Care [MH] OR Physicians, Family [MH] OR general practice [tiab] OR general practitioner [tiab] OR primary care OR "Physical Therapy Modalities"[MH] OR "Physical Therapy Specialty"[MH] OR Rehabilitation [MH] OR physiotherapy*[tiab]OR physical therapy [tiab] OR physical therapist* [tiab] OR physical therapeutic [tiab]

#4-1 MT Pillastrini (slightly adapted)
Chiropractic [MH] OR Manipulation, Osteopathic [MH] OR Musculoskeletal Manipulations [MH] OR Joint Mobilization* OR Manipulative OR Manual Therap* OR "Muscle Strengthening" OR "Muscle Stretching" OR Myofascial* OR Osteopathic Manipulation* OR "Proprioceptive Neuromuscular Facilitation" OR Spinal Manipulation* OR "Static Stretching" OR Trigger Point* OR Exercise Movement Techniques [MH] OR Exercise Therapy [MH] OR Manipulation, Orthopedic [MH] OR Massage [MH] OR Muscle Relaxation [MH] OR Muscle Stretching Exercises [MH] OR Osteopathic Medicine [MH] OR Traction [MH] OR "Clinical Reasoning" OR "Exercise Therapy" OR "Joint Range of Motion" OR Joint Stabilization* OR Joint Stabilisation* OR Manipulation* OR Manual Intervention* OR "Massage" OR Mobilization* OR Mobilisation OR Motor Control* OR "Motor Learning" OR "Muscle Relaxation" OR "Muscle Strength Training" OR Neurodynamic* OR "Orthopedic Manipulation" OR Osteopathic* OR "Osteopathic Medicine" OR "Passive Range of Motion" OR "Passive Stretching" OR Postural OR Postural Adjustment* OR "Postural Balance" OR "Postural Control" OR "Postural Stability" OR "Range of Motion" OR Stabilization* OR Stretching OR Thrust* OR Traction OR manual medicine* [tiab]

(P) #5 neck pain
Neck Pain OR non-specific neck pain OR neck complaints OR "Neck Injuries"[Mesh] OR neck injury OR Whiplash OR WAD OR Whiplash-associated disorders OR Cervical Radiculopathy OR neck shoulder OR neck arm OR "Thoracic Outlet Syndrome"[Mesh] OR neck pain with radiculopathy OR Cervical degenerative disc disease OR cervical disc disease OR cervical spondylosis

Embase:
(I) #1 multivariable prognostic models van Oort: 'decision support system'/exp OR 'decision support system' OR 'predictive value'/exp OR 'predictive value' OR 'clinical prediction':ab,ti OR prognos*:ab,ti OR predict*:ab,ti

(I) #4 primary musculoskeletal care
'physiotherapy'/exp OR physiotherap*:ab,ti OR 'physical therapy':ab,ti OR 'physical therapist':ab,ti OR 'physical therapists':ab,ti OR 'community based rehabilitation'/exp OR 'community based rehabilitation' OR 'functional assessment'/exp OR 'functional assessment' OR 'functional training'/exp OR 'functional training' OR 'geriatric rehabilitation'/exp OR 'geriatric rehabilitation' OR 'home rehabilitation'/exp OR 'home rehabilitation' OR 'muscle training'/exp OR 'muscle training' OR 'occupational therapy'/exp OR 'occupational therapy' OR 'primary health care'/exp OR 'primary health care' OR 'general practitioner'/exp OR 'general practitioner' OR 'general practice'/exp OR 'general practice' OR 'primary care':ab,ti OR 'general practice':ab,ti OR 'general practitioner':ab,ti

#4-1 MT Pillastrini (slightly adapted)
('chiropractic'/exp OR 'chiropractic' OR 'manipulation, osteopathic'/exp OR 'manipulation, osteopathic' OR manipulative OR ('medicine'/exp OR medicine) OR 'joint mobilization'/exp OR 'joint mobilization' OR 'manipulative' OR 'muscle strengthening'/exp OR 'muscle strengthening' OR 'muscle stretching'/exp OR 'muscle stretching' OR myofascial*:ab,ti OR

'proprioceptive neuromuscular facilitation' OR 'static stretching' OR 'trigger point'/exp OR 'trigger point' OR 'exercise movement techniques'/exp OR 'exercise movement techniques' OR 'manipulation, orthopedic'/exp OR 'manipulation, orthopedic' OR 'muscle stretching exercises'/exp OR 'muscle stretching exercises' OR 'traction'/exp OR 'traction' OR 'clinical reasoning' OR 'exercise therapy'/exp OR 'exercise therapy' OR 'joint range of motion' OR 'joint stabilization' OR manipulation*:ab,ti OR 'manual intervention' OR 'massage'/exp OR 'massage' OR mobilization*:ab,ti OR 'motor control'/exp OR 'motor control' OR 'motor learning' OR 'muscle relaxation'/exp OR 'muscle relaxation' OR 'muscle strength training' OR neurodynamic*:ab,ti OR 'orthopedic manipulation'/exp OR 'orthopedic manipulation' OR osteopathic*:ab,ti OR 'osteopathic medicine'/exp OR 'osteopathic medicine' OR 'passive range of motion' OR 'passive stretching' OR 'physical therapy'/exp OR 'physical therapy' OR 'physiotherapy'/exp OR 'physiotherapy' OR 'postural' OR adjustment*:ab,ti OR 'postural balance'/exp OR 'postural balance' OR 'postural control' OR 'postural stability' OR 'range of motion'/exp OR 'range of motion' OR 'reflexology'/exp OR 'reflexology' OR stabilization*:ab,ti OR 'stretching'/exp OR 'stretching' OR thrust*:ab,ti OR 'physical medicine'/exp OR 'physical medicine')

(P) #5 neck pain
('neck pain'/exp OR 'neck pain' OR 'non-specific neck pain' OR 'neck complaints' OR 'neck injuries'/exp OR 'neck injuries' OR 'neck injury'/exp OR 'neck injury' OR 'whiplash injury'/exp OR 'whiplash injury' OR 'wad' OR 'whiplash-associated disorders' OR 'cervical radiculopathy'/exp OR 'cervical radiculopathy' OR 'cervicobrachial neuralgia'/exp OR 'cervicobrachial neuralgia' OR 'neck shoulder' OR 'neck arm' OR 'thorax outlet syndrome'/exp OR 'thorax outlet syndrome' OR 'neck pain with radiculopathy' OR 'cervical degenerative disc disease' OR 'cervical disc disease' OR 'cervical spondylosis/exp')

Cinahl:
(I) #1 multivariable prognostic models van Oort:
((MM "Predictive Research") or (MM "Predictive Validity") or (MM "Predictive Value of Tests") or (MM "Decision Support Systems, Clinical") or (MM "Prognosis+"))

(I) #4 primary musculoskeletal care FT van Oort:
((MM "Rehabilitation+") or (MM "Physical Therapy+") or (MM "Pediatric Physical Therapy") or (MM "Research, Physical Therapy") or (MM "Physical Therapists") or ("Rehabilitation+") or ("Physical Therapy+") or ("Pediatric Physical Therapy") or ("Research, Physical Therapy") or ("Physical Therapists")) or ("Primary Health Care") OR ("Physicians, Family") OR ("general practice") OR ("general practitioner") OR ("primary care") OR ("Physical Therapy Modalities") OR ("Physical Therapy Specialty") OR ("Rehabilitation") OR ("physiotherapy+") OR ("physical therapy") OR ("physical therapist") OR ("physical therapeutic"))

#4-1 MT Pillastrini (slightly adapted)
(MH "Manual Therapy+") or (MH "Chiropractic+") or (MH "Osteopathic Medicine") OR (MH "Osteopathy+") OR (MH "Manipulation, Osteopathic") or (MH "Manipulation, Orthopedic") or ("Manual Therapy+") or ("Chiropractic+") or ("Osteopathic Medicine") OR ("Osteopathy+") OR ("Manipulation, Osteopathic") or ("Manipulation, Orthopedic") OR

manipulative OR “joint mobilization” OR “muscle strengthening” OR “muscle stretching” OR “myofascial+”:ab,ti OR “proprioceptive neuromuscular facilitation” OR “static stretching” OR “trigger point” OR “exercise movement techniques” OR “manipulation, orthopedic” OR “muscle stretching exercises” OR “traction” OR “clinical reasoning” OR “exercise therapy” OR “joint range of motion” OR “joint stabilization” OR “manipulation+” OR “manual intervention” OR “massage” OR “mobilization+” OR “motor control” OR “motor learning” OR “muscle relaxation” OR “muscle strength training” OR “neurodynamic+” OR “orthopedic manipulation” OR “osteopathic+” OR “osteopathic medicine” OR “passive range of motion” OR “passive stretching” OR “postural” OR “adjustment+” OR “postural balance” OR “postural control” OR “postural stability” OR “range of motion” OR ‘reflexology’/exp OR “reflexology” OR “stabilization+” OR “stretching” OR “thrust+” OR “physical medicine”

(P) #5 neck pain

“neck pain” or “neck complaints” or “non-specific neck pain” or “neck injuries” or whiplash injury” or whiplash-associated disorders OR cervical radiculopathy or cervicobrachial or “neck shoulder” or “ neck arm” or thorax outlet syndrome” or cervical spondylosis (MH “Neck Pain”) or (MH “Neck Injuries+”) or (MH “Whiplash Injuries”) or (MH “Cervical Plexus+”) or (MH “Thoracic Outlet Syndrome”) or (MH “Osteoarthritis, Cervical”) or (“neck pain”) or (“Neck Injuries+”) or (“Whiplash Injuries”)

Appendix 2. Study characteristics derivation only studies

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
Angst et al 2014 [55] (and rectification) Swiss inpatient clinic	WAD-chronic. n=185, mean age 37.4 (SD 11.7), 79,4% female, disease duration 13.3 months (SD 10.7). Interdisciplinary program.	Pain (NASS), 6 months	Pain (NASS); Function (NASS): change; Anxiety (HADS), sports; working capacity; Catastrophizing (CSQ): change; Smoking	R2=72,1%	Univariate variables: 22 Multivariate variables: 21 Predictors in model: 8 Analysed: 103/185 Participants/Predictors: 103/21=5
		Bodily pain (SF-36), discharge	Bodily pain (SF-36); Pain decrease (CSQ): baseline; Pain decrease (CSQ): change; Depression (HADS): change; Physical function (SF-36): change; Depression (HADS); Physical function (SF-36); Social functioning (SF-36): change; Smoking; Sports	R2=54,6%	Univariate variables: 22 Multivariate variables: 21 Predictors in model: 10 Analysed: 175/185 Participants/Predictors: 175/21=8
		Function (SF-36 physical functioning), 6 months	Depression (HADS): change; Physical function (SF-36); Depression (HADS); Bodily pain (SF-36): change; Bodily pain (SF-36): baseline; Age; Sports	R2=63,4%	Univariate variables: 22 Multivariate variables: 21 Predictors in model: 7 Analysed: 103/185 Participants/Predictors: 103/21=5
		Function (NASS), discharge	CSQ Catastrophizing change; NASS Function baseline; NASS Pain change; NASS Pain baseline; HADS Depression baseline; HADS Depression change; Sex	R2=53,3%	Univariate variables: 22 Multivariate variables: 21 Predictors in model: 7 Analysed: 175/185 Participants/Predictors: 175/21=8
Asenlof et al 2013 [67]; Sweden, emergency dept., 2007-2009	WAD- acute, grade 1 and 2. n=98, mean age 34.4 (SD 11.4), 53,1 % female, WAD1 l 49.0%, WAD 2 51%, median NPRS 2/10. No need for further treatment	Function (PDI), 12 months	Function (PDI)	Adjusted R2=66%	Univariate variables: 6 Multivariate variables: 6 Predictors in model: 1 Analysed: 73/98 Participants/Predictors: 73/12=6
Atherton et al 2006 [61]; UK emergency dept., 2002 - 2003	WAD-acute. n=765 full baseline data. Median age 34 years (IQR 25–44 years), 56% female, 75% WAD 1.	Pain, 12 months	Pre-collision widespread pain; Vehicle type; number of WAD symptoms; Function (NDI); Psychological distress (GHQ)	No information	Univariate variables: 26+31 category Df Multivariate variables: 7 Predictors in model: 5 Analysed: 480/765 Events/Non-events: 128/352 EPV: 128/64=2
Baltov et al 2008 [68]; Canada, Hospital Rehab	WAD-chronic. n=28, mean age 33.29 y (SD8.96), 64.3% female, mean NDI 22.89 (SD8.5), all on sick leave. Individualized	Function (NDI), discharge	Function (NDI)	Adjusted R2=61,3%	Univariate variables: 21 Multivariate variables: 3 Predictors in model: 1 Analysed: 25/28 Participants/Predictors: 25/24=1
		Function (NDI), 3 months	Function (NDI)	Adjusted R2=59,8%	Univariate variables: 21 Multivariate variables: 3 Predictors in model: 1

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
	multidisciplinary rehabilitation				Analysed: 23/28 Participants/Predictors: 23/24=1
Bohman et al 2012 [60]; Canada, insurance company 1997-1999*	WAD-acute. n=599, mean age 39 (SD 15), 69,3% female, mean baseline neck pain intensity of 6.8/10 (SD 2.0).	Recovery (GROC), 6 months	Age; Days to collision reporting; Neck pain intensity; Low back pain intensity; Other pain; Pre-collision headache; Recovery expectations.	C-index 0.68 (0.65-0.71) C-index internal validation 0.67 (90.63-0.70)	Univariate variables: 25+33 category Df Multivariate variables: 22 Predictors in model: 7 Analysed: 633/680 Events/Non-events: 484/115 EPV: 115/80=1.4
Buitenhuis et al 2006 [82]; Dutch insurance company	WAD-acute. n=240, mean age 36.0 (SD 12.8), 63,7% female	Severity (11-items severity score) with PTSD clustered present, 6 months	Gender; Neck pain; Hyper arousal symptoms	No information	Univariate variables: 22+10 category Df Multivariate variables: >5 Predictors in model: 3 Analysed: 79/134 Events/Non-events: 79/55 EPV: 55/>37=<1.5
		Severity (11-items severity score) with PTSD categorical present, 6 months	Gender; Neck pain; Dizziness.	No information	Univariate variables: 24+10 category Df Multivariate variables: >6 Predictors in model: 3 Analysed: 79/134 Events/Non-events: 79/55 EPV: 55/>40=<1.4
		Severity (11-items severity score) with PTSD clustered present, 12 months	Neck pain; Hyper arousal symptoms	No information	Univariate variables: 22+10 category Df Multivariate variables: >5 Predictors in model: 2 Analysed: 62/134 Events/Non-events: 62/72 EPV: 62/>37=<1.7
		Severity (11-items severity score) with PTSD categorical present, 12 months	Gender; Dizziness.	No information	Univariate variables: 24+10 category Df Multivariate variables: >6 Predictors in model: 2 Analysed: 62/134 Events/Non-events: 62/72 EPV: 62/>40=<1.6
Bunketorp et al 2006 [83]; Swedish WAD rehabilitation centre.	WAD-subacute. n=47, mean age 31 y, female 64%	Function (PDI), max. 3 months	Self-Efficacy Scale	Adjusted R2=42%	Univariate variables: 7 Multivariate variables: 9 Predictors in model: 1 Analysed: 40/47 Participants/Predictors: 40/16=2.5
Cai et al 2011 [46]; Singapore outpatient hospital physical therapy clinic.	Neck pain. n=103 ;mean age 48,8 years, 37,9 % female, mean duration 30,6 weeks. Traction.	Composite endpoint (NPRS, NDI, GROC). Two weeks?	Pain intensity (NRS); FABQ Work; Response traction test; Pain below shoulder level	R2= 38% H-L statistic p=0.77	Univariate variables: 40 Multivariate variables: 10 Predictors in model: 4 Analysed: 103/103 Events/Non-events: 47/50 EPV: 47/50=0.9
Carstensen et al 2015 [63]; Denmark emergency dept. or family physicians, 2001-2003	WAD-acute. n=719, mean age 34.4 years, 64.4% female	Pain (VAS), 12 months	Pre-collision sickness benefit; Pre-collision pain condition; Gender; Inclusion neck pain	H-L statistic p=0.17 AUC: 0.80	Univariate variables: 9+8 category Df Multivariate variables: 8 Predictors in model: 4 Analysed: 476/719 Events/Non-events: 167/309 EPV: 167/25=6.7
Cecchi et al 2011 [84]; Italy	Neck pain-chronic, non-	Pain (NPQ), discharge	Neck pain related use of drugs.	R2=20%	Univariate variables: 23 Multivariate variables: 22

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
outpatient rehabilitation clinic 2008-2009	specific (no WAD). n=178, mean age 65 (SD12.5), 77% female, NPQ 40.7 (SD17.1). Standardized exercise PT protocol and advise				Predictors in model: 1 Analysed: 162/178 Events/Non-events: 73/89 EPV: 73/45=1.6
		Pain (NPQ), 12 months	Neck pain related use of drugs; catastrophizing (PCS)	R2=16%	Univariate variables: 23 Multivariate variables: 22 Predictors in model: 2 Analysed: 162/178 Events/Non-events: 90/72 EPV: 72/45=1.6
Chiarotto et al 2015 [56]; Italy, physiotherapy clinics, 2012-2013	WAD, grade 2 and 3. n=39, mean age 41.1 (SD, 11.9), 51% female, mean NDI 28.6 (SD10.5). Tailored multimodal manual therapy	Pain (NRS), 1 day	Pain intensity previous week, Pain catastrophizing (PCS)	R2=36%	Univariate variables: 12+8 category Df Multivariate variables: 11 Predictors in model: 2 Analysed: 37/39 Participants/Predictors: 37/31=1.2
		Pain (NRS), 1 week	Pain intensity previous week, Pain catastrophizing (PCS)	R2=35%	Univariate variables: 12+8 category Df Multivariate variables: 11 Predictors in model: 2 Analysed: 37/39 Participants/Predictors: 37/31=1.2
		Function (NDI), immediate after therapy	Disability (NDI); Pain catastrophizing (PCS)	R2=49%	Univariate variables: 12+8 category Df Multivariate variables: 11 Predictors in model: 2 Analysed: 37/39 Participants/Predictors: 37/31=1.2
Cleland et al 2007 [11]; USA physical therapy clinic, 2004-2005	Neck pain with or without unilateral upper extremity symptoms. n=80, mean age 42 (SD 11,3), 68% female, mean duration 80 (SD 70,6) days, NDI 34,9 (SD 10,1). Thoracic manipulations, cervical ROM exercise, education	Recovery (GROC), 2nd or 3rd session	Duration <30 days; No symptoms distal to shoulder; Looking upwards do not aggravate; FABQPA <12; Diminished T3–T5 kyphosis; Cervical extension <30 degrees.	R2=68%	Univariate variables: 34 Multivariate variables: 10 Predictors in model: 6 Analysed: 78/80 Events/Non-events: 42/36 EPV: 36/44=0.8
Cleland et al 2007 [85]; USA physiotherapy clinics 2004-2006	Neck pain-radiculopathy or neck-arm pain. n=101, mean age 50,8 yrs, 64% female, mean duration 60,2 days, median NDI 28,6. Individual PT intervention	Recovery (if all outcome measures are met of: GROC, NDI, PSFS, NPRS), discharge or last examination	Age <54; Dominant arm is not affected; Looking down not worsen; Multimodal treatment	R2=45%	Univariate variables: 23 Multivariate variables: 8 Predictors in model: 4 Analysed: 96/101 Events/Non-events: 50/46 EPV: 46/31=1.5
Cobo et al 2010 [42]; orthopaedic rehab dept. Spain, 2005-2007	WAD-acute, grade 1 and 2. n=682, mean age 35.6 (SD13.5), 66,8% female,	Pain (VAS), 6 months	Pain (VAS); Age; Pain (NPH), Dizziness; Self-employed	Adjusted R2=20%	Univariate variables: 36+10 category Df Multivariate variables: 15 Predictors in model: 5 Analysed: 557/682 Participants/Predictors:

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
	45,4% severe pain, 41,5% labour disability. Hospital rehab received.				557/51=10.9
Dagfinrud et al 2013 [69]; Norway, manual therapy clinics	Neck pain. n=81, mean age 43.4 (SD14.4), 72% female, NDI 27.13 (SD13.1). Individualised manual therapy treatment.	Function (NDI), 8 weeks	Age; Gender; Function (NDI), Pain duration; Clinician's prediction, ÖMPQ	Adjusted R2=64%	Univariate variables: 13+13 category Df Multivariate variables: 7 Predictors in model: 6 Analysed: 81/81 Participants/Predictors: 81/19=4.3
Gun et al 2005 [86]; Australia, emergency depts., medical-, physiotherapy clinics	WAD-acute. n=147, mean age 35.6 (SD14,7), 73% female, neck pain (5,9 SD 2,5). Various therapy.	Function (NPOS), 12 months	Age; SF-36 four subscales (Mental, Physical, Bodily pain, Role emotional); Head rest; Lawyer consult; Claim; Treated; Vehicle not drivable	No information	Univariate variables: 20 Multivariate variables: 10 Predictors in model: 5 Analysed: 135/147 Participants/Predictors: 135/>30=<4.5
		Pain (VAS), 12 months	Age; SF-36 four subscales (Mental, Physical, Bodily pain, Role emotional); Head rest; Lawyer consult; Claim; Treated; Vehicle not drivable	No information	Univariate variables: 20 Multivariate variables: 10 Predictors in model: 5 Analysed: 135/147 Participants/Predictors: 135/>30=<4.5
Hanney et al 2013 [47]; USA, physical therapy, 2009-2011	Neck pain. n=91, mean age 45,7 (SD 13,3), 75,8% female, NDI 17,7/50. (SD 7,9). Multimodal PT program (stretching, exercises)	Recovery (GROC), 6 weeks	NDI score < 18/50; Shoulder protraction; Patient does not cycle; Cervical side bending < 32°; FABQ-Physical Activity < 15.	R2=33% H-L statistic p=0.58	Univariate variables: 43 Multivariate variables: 7 Predictors in model: 5 Analysed: 91/91 Events/Non-events: 50/41 EPV: 41/50=0.8
Hartling et al 2002 [87]; Canada, emergency departments, 1995-1998	WAD-acute. n=353; 65% female	Pain (Severity/Frequency scale), 6 months	Number of symptoms; Age group;	No information	Univariate variables: 44+55 category Df Multivariate variables: 99-8=91 Predictors in model: 2 Analysed: 334/353 Events/Non-events: 118/216 EPV: 118/91=1.3
		Pain (Severity/Frequency scale), 6 months	Upper back pain; Upper extremity numbness or weakness; Vision disturbances; Age group	No information	Univariate variables: 44+55 category Df Multivariate variables: 99-8=91 Predictors in model: 2 Analysed: 334/353 Events/Non-events: 118/216 EPV: 118/91=1.3
Hendriks et al 2005 [62]; Dutch GP and emergency dept., 1999-2002	WAD-acute. n=125, mean age 34.1 (SD 10.1); 61% female, mean neck pain intensity 42.1 (SD 25.5).	Recovery (VAS-pain or VAS-activities without pain medication), 4 weeks	Pain (VAS); Somatisation (SCL-90 subscale); Sleep difficulties (SCL-90 subscale); Work disability (VAS)	R2=65.4% AUC=0.93 (0.88-0.97)	Univariate variables: 36+3 category Df Multivariate variables: 7 Predictors in model: 4 Analysed: 125/125 Events/Non-events: 80/45 EPV: 45/46=1

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
	Treatment by GP (education, advise) and PT (education, advise, graded activity, exercise).	Recovery (VAS-pain or VAS-activities without pain medication), 12 weeks	Female; Unprepared for collision; Pain (VAS); Sleep difficulties (SCL-90 subscale); Work disability (VAS)	R2=52.6% AUC=0.88 (0.82-0.94)	Univariate variables: 36+3 category Df Multivariate variables: 10 Predictors in model: 5 Analysed: 125/125 Events/Non-events: 49/76 EPV: 49/49=1
		Recovery (VAS-pain or VAS-activities without pain medication), 12 months	Female; Low level education; Pain (VAS); Somatisation (SCL-90 subscale); Work disability (VAS)	R2=45.9%; AUC=0.86 (0.80-0.92)	Univariate variables: 36+3 category Df Multivariate variables: 11 Predictors in model: 5 Analysed: 119/125 Events/Non-events: 39-45/74-80 EPV: 45/50=0.9
Hill et al 2007 [48]; UK GP referrals to physical therapy, 2000-2002	Neck pain, non-specific. n=350, mean age 51 years (range 23-84); NPQ mean 37.2 (SD 14.0) 63% female. Three physical therapy treatments (advice and exercise alone, or in addition to manual therapy or pulsed short wave diathermy)	Recovery (GROC), 6 weeks	Manual occupation	R2=14%	Univariate variables: 16+12 category Df Multivariate variables: 15+12 category Df Predictors in model: 1 Analysed: 316/346 Events/Non-events: 103/213 EPV: 103/27=3.8
		Pain (NPQ), 6 weeks	Manual occupation, Lower physical health (SF12-physical component summary)	R2=14%	Univariate variables: 16+12 category Df Multivariate variables: 15+12 category Df Predictors in model: 2 Analysed: 316/346 Events/Non-events: 158/158 EPV: 158/27=5.9
		Recovery (GROC), 6 months	Age category ≥60 years; Expectations; Comorbid low back pain; Catastrophizing (PCS); Higher pain (NPQ)	R2=35%	Univariate variables: 16+12 category Df Multivariate variables: 15+12 category Df Predictors in model: 6 Analysed: 321/346 Events/Non-events: 125/196 EPV: 125/27=4.6
		Pain (NPQ), 6 months	Catastrophizing (PCS)	R2=17%	Univariate variables: 16+12 category Df Multivariate variables: 15+12 category Df Predictors in model: 1 Analysed: 321/346 Events/Non-events: 140/181 EPV: 140/27=5.2
Hoving et al 2004 [57]; Dutch, General practices, 1997-1998	Neck pain. n=183. Age ≥40 years 66.7% female, mean NDI 14.5/50 (SD 7.0). Continued GP care, physical therapy or manual therapy	Recovery (GROC), 7 weeks	Age ≥40 y; Headache	No information	Univariate variables: 11 Multivariate variables: 14 Predictors in model: 2 Analysed: 183/183 Events/Non-events: 94/89 EPV: 89/14=6.4
		Pain (NRS), 7 weeks	Age ≥40 y; Headache; Low back pain; Pain intensity.	Adjusted R2=24%	Univariate variables: 11 Multivariate variables: 14 Predictors in model: 4 Analysed: 183/183 Events/Non-events: 94/89 EPV: 89/14=6.4

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
		Function (NDI), 7 weeks	Age ≥40 years; Neck function; Low back pain	Adjusted R2=36%	Univariate variables: 11 Multivariate variables: 14 Predictors in model: 3 Analysed: 183/183 Events/Non-events: 94/89 EPV: 89/14=6.4
		Recovery (GROC), 12 months	Age ≥40 years; Previous trauma; Low back pain; No 2 wk change in neck pain; High severity of physical dysfunctioning.	No information	Univariate variables: 11 Multivariate variables: 14 Predictors in model: 5 Analysed: 178/183 Events/Non-events: 113/65 EPV: 65/14=4.6
		Pain (NRS), 12 months	Age ≥40 years; Previous episode; Low back pain; ≥13 wk. duration; Pain intensity	Adjusted R2=30%	Univariate variables: 11 Multivariate variables: 14 Predictors in model: 5 Analysed: 178/183 Events/Non-events: 113/65 EPV: 65/14=4.6
		Function (NDI), 12 months	Age ≥40 years; Neck function; ≥13 wk. duration; Traumatic; No 2wk change in neck pain; Low back pain	Adjusted R2=26%	Univariate variables: 11 Multivariate variables: 14 Predictors in model: 6 Analysed: 178/183 Events/Non-events: 113/65 EPV: 65/14=4.6
Kjellman et al 2002 [88]; Sweden physiotherapy and chiropractic clinics, 1993-1997	Neck pain-with and without radiation. n=193, mean age 39.3 (SD 10.3), 76% female, mean pain 49.7 (SD 22.6). Received primary physiotherapy or chiropractic care.	Function (Oswestry), 12 months	Pain intensity (VAS); Well-being; Expectations treatment; Duration current episode.	Adjusted R2=32%	Univariate variables: 18+28 category Df Multivariate variables: 15+3 category Df Predictors in model: 4 Analysed: 156/193 Participants/Predictors: 156/18=8.7
		Pain (VAS), 12 months	Oswestry; Duration current episode; Similar problems	Adjusted R2=24%	Univariate variables: 18+28 category Df Multivariate variables: 15+3 category Df Predictors in model: 3 Analysed: 156/193 Participants/Predictors: 156/18=8.7
Kyhlbäck et al 2002 [43], Sweden, orthopaedic clinic, 1997-1998	WAD-acute. n=98, mean age 35, 66% female	Pain (VAS), 3 months	Self-efficacy scale	R2=14.5%	Univariate variables: 6 Multivariate variables: 6 Predictors in model: 1 Analysed: 68/98 Participants/Predictors: 68/6=11.3
		Function (PDI), 3 months	Self-efficacy scale, Age	R2=47.5%	Univariate variables: 6 Multivariate variables: 6 Predictors in model: 2 Analysed: 76/98 Participants/Predictors: 76/6=12.7
		Pain (VAS), 12 months	Self-efficacy scale, WAD grade; Gender	R2=23.6%	Univariate variables: 6 Multivariate variables: 6 Predictors in model: 3 Analysed: 70/98 Participants/Predictors: 70/6=11.7
		Function (PDI), 12 months	Self-efficacy scale, Age; Gender	R2=36.3%	Univariate variables: 6 Multivariate variables: 6 Predictors in model: 3 Analysed: 78/98 Participants/Predictors:

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
					78/6=13
Landers et al 2008 [49]; physiotherapy clinics	Neck pain. n=79 mean age 49.6 y (SD 12.7), 71% female. Personalized non-protocol PT care.	Function (NDI), 12 weeks	Function (NDI); FABQ-PA; Cervical non organic signs	R2=67.5%	Univariate variables: 5 Multivariate variables: 7+2 category Df Predictors in model: 3 Analysed: 79/79 Events/Non-events: 29/50 EPV: 29/9=3.2
Lankester et al 2006 [44]; UK medio Legal reports.	WAD. n=277; mean age 39.9 y (range 15–81),	Function (NDI), 9 months to 5 years; model with pre-accident predictors	Pre-existing back pain; Known psychological / anxiety disorder; Frequent GP attendance	No information	Univariate variables: 14+7 category Df Multivariate variables: 6+5 category Df Predictors in model: 3 Analysed: 176/277 Participants/Predictors: 176/11=16
		Function (NDI), 9 months to 5 years; model with accident predictors	Front position in vehicle	No information	Univariate variables: 14+7 category Df Multivariate variables: 3+2 category Df Predictors in model: 1 Analysed: 277/277 Participants/Predictors: 277/5=55.4
		Function NDI, 9 months to 5 years; model with response predictors	Early onset of symptoms; Pain radiating away from the neck.	No information	Univariate variables: 14+7 category Df Multivariate variables: 3 Predictors in model: 2 Analysed: 277/277 Participants/Predictors: 277/3=92.3
		Symptoms severity (GBG), 9 months to 5 years model with pre-accident predictors	Pre-existing back pain; Known psychological / anxiety disorder; Frequent GP attendance	No information	Univariate variables: 14+7 category Df Multivariate variables: 6+5 category Df Predictors in model: 3 Analysed: 176/277 Participants/Predictors: 176/11=16
		Symptoms severity (GBG), 9 months to 5 years; model with accident predictors	Front position in vehicle	No information	Univariate variables: 14+7 category Df Multivariate variables: 3+2 category Df Predictors in model: 1 Analysed: 277/277 Participants/Predictors: 277/5=55.4
		Symptoms severity (GBG), 9 months to 5 years; model with response predictors	Early onset of symptoms; Pain radiating away from the neck; Abnormal neurological finding.	No information	Univariate variables: 14+7 category Df Multivariate variables: 3 Predictors in model: 2 Analysed: 277/277 Participants/Predictors: 277/3=92.3
Michaelson et al 2004 [89]; Sweden rehab centre 1997-1999	Neck pain-chronic. n=136, mean age 42, 73% female, mean duration 108 (SD87) months, pain intensity 60 (SD17). Multimodal program (physical + cognitive behavioural)	Pain (VAS), 4 weeks	Optimism index; Sociability index; Endurance index; Average pain intensity.	R2=42%	Univariate variables: 17 Multivariate variables: 17+2 category Df Predictors in model: 4 Analysed: 131/136 Events/Non-events: 57/79 EPV: 57/19=3
		Pain (VAS), 12 months	Optimism index; Sociability scale; Age; Other symptoms-index; Average pain intensity.		Univariate variables: 17 Multivariate variables: 17+2 category Df Predictors in model: 5 Analysed: 106/136 Events/Non-events: 32/74 EPV: 32/19=1.7

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
Nederhand et al 2004 [64]; Dutch emergency centre 1999-2001	WAD -acute; grade 1,2. n=90; 18-70 y	Function (NDI), 24 weeks; model with VAS predictor	VAS	AUC=0.71 (0.57–0.87)	Univariate variables: 10+2 category Df Multivariate variables: 2 Predictors in model: 1 Analysed: 82/90 Events/Non-events: 27/55 EPV: 27/14=1.9
		Function (NDI), 24 weeks; model with TSK predictor	TSK	AUC=0.77 (0.63–0.91)	Univariate variables: 10+2 category Df Multivariate variables: 2 Predictors in model: 1 Analysed: 82/90 Events/Non-events: 27/55 EPV: 27/14=1.9
		Function (NDI), 24 weeks NDI; model with pain cognition	Pain Cognition List–Experimental version	AUC=0.73 (0.59–0.88)	Univariate variables: 10+2 category Df Multivariate variables: 2 Predictors in model: 1 Analysed: 82/90 Events/Non-events: 27/55 EPV: 27/14=1.9
		Function (NDI), 24 weeks; model with isometric muscle Activity	Isometric Muscle Activity	AUC=0.68 (0.52–0.84)	Univariate variables: 10+2 category Df Multivariate variables: 2 Predictors in model: 1 Analysed: 82/90 Events/Non-events: 27/55 EPV: 27/14=1.9
		Function (NDI), 24 weeks	NDI; TSK	R2=42%	Univariate variables: 10+2 category Df Multivariate variables: 2 Predictors in model: 2 Analysed: 82/90 Events/Non-events: 27/55 EPV: 27/14=1.9
Nee et al 2013 [65]; Australian general community and GP practices	Neck pain and nerve related arm pain. n=40, mean age y 47 (SD 8), 65% Female, NDI 12.7/50 (SD4.2). Manual therapy and neurodynamic treatment.	Recovery (GROC), 3-4 weeks	S-LANSS; Age; ULNT-median	R2=46% AUC= 0.85 (0.72-0.98) H-L statistic not enough power	Univariate variables: 43+2 category Df Multivariate variables: 38 Predictors in model: 3 Analysed: 38/40 Events/Non-events: 19/21 EPV: 19/28=0.5
Nieto et al 2013 [90]; Rehabilitation centres, Spain 2006-2007	WAD (within 3 months). n=147; mean age 34,8 (SD 10.15), 75,6% female, NDI 37.8 (SD 15,17). Rehab programme	Function (NDI), 6 months	Fear of movement (TSK), Pain (NRS), Function (NDI)	R2=48%	Univariate variables: 8+8 category Df Multivariate variables: 8 Predictors in model: 3 Analysed: 123/147 Participants/Predictors: 123/16=7.7
		Pain (NRS), 6 months	Pain (NRS), Function (NDI)	R2=31%	Univariate variables: 8+8 category Df Multivariate variables: 8 Predictors in model: 2 Analysed: 123/147 Participants/Predictors: 123/16=7.7
Pape et al 2007 [91]; Insurance	WAD-acute. n=1310 >16 yr.	Pain (Severity/Frequency), 3 years	Direction collision; Memory/concentration, Neck and/or shoulder	ROC curve	Univariate variables: 80+93 category Df Multivariate variables: 16

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
company liability claims, Norway 1996-1997			pain before the accident; Difficulties bodily tension; Difficulties climb stairs; Difficulties bending forward; Difficulties heavy labour; Perception ability to work in half a year's time.		Predictors in model: 8 Analysed: 636/1310 Events/Non-events: 97/549 EPV: 87/189=0.5
Peterson et al 2012 [66]; multiple chiropractic practices, Switzerland	Neck pain-acute and neck pain-chronic. n=529; Acute n=274: mean age 40 (SD 12,58), 59,1 % female, BQ 33,96 (SD 15,26); Chronic n=255: 41,8 (SD 13,87), 65,1% female, BQ 30,50 (SD 14,18). Chiropractic treatment.	Recovery (PGIC), 1 month; acute model	Depression (BNQ-subscale); Pain change to 1 week (BNQ-subscale); Pain change to 1 week (NRS); Recovery at 1 week (PGIC)	Adjusted R2=21.7 AUC=0.79 (0.70-0.88)	Univariate variables: 31 Multivariate variables: 6 Predictors in model: 4 Analysed: 180/215 Events/Non-events: 237/37 EPV: 37/37=1
		Recovery (PGIC), 3 months acute model	Function change to 1 month (NBQ); Recovery at 1 week (PGIC)	Adjusted R2=28.8 AUC=0.82 (0.72-0.92)	Univariate variables: 42 Multivariate variables: 16 Predictors in model: 2 Analysed: 146/197 Events/Non-events: 213/43 EPV: 43/58=0.7
		Recovery (PGIC), 1 month; chronic model	Recovery at 1 week (PGIC)	Adjusted R2=12.7 AUC=0.66 (0.57-0.75)	Univariate variables: 31 Multivariate variables: 2 Predictors in model: 1 Analysed: 156/204 Events/Non-events: 159/96 EPV: 96/33=2.9
		Recovery (PGIC), 3 months; chronic model	Recovery at 1 month (PGIC)	Adjusted R2=19,9 AUC=0.71 (0.62-0.79)	Univariate variables: 42 Multivariate variables: 9 Predictors in model: 1 Analysed: 133/185 Events/Non-events: 179/76 EPV: 76/51=1.5
Pool et al 2010 [92]; Dutch physical/ manual therapy centres, 2003-2004	Neck pain-subacute. n=146, mean age 45.1 (SD 11.2), 61% female, NDI 14.0 (SD 6.8). Manual therapy treatment and behavioural graded activity programme	Recovery (GPE), 12 weeks	Headache; Preference for physical therapy.	R2=17%	Univariate variables: 21+5 category Df Multivariate variables: 21 Predictors in model: 2 Analysed: 146/146 Events/Non-events: 103/43 EPV: 43/47=0.9
		Recovery (GPE) 12 months	Less fear of movement (TSK)	R2=6%	Univariate variables: 21+5 category Df Multivariate variables: 21 Predictors in model: 1 Analysed: 136/146 Events/Non-events: 105/31 EPV: 31/47=0.7
		Pain (NRS), 12 weeks	Fear of movement (TSK); Male gender; Severity of complaints	R2=16%	Univariate variables: 21+5 category Df Multivariate variables: 21 Predictors in model: 3 Analysed: 146/146 Events/Non-events: 71/75 EPV: 71/47=1.5
		Function (NDI), 12 weeks	Fear of movement (TSK); Somatisation	R2=30%	Univariate variables: 21+5 category Df

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
			(4DSQ), male; Age; Level of chronicity (GCPS); Internal pain control (PCCL)		Multivariate variables: 21 Predictors in model: 3 Analysed: 146/146 Participants/Predictors: 146/47=3.1
		Function (NDI), 12 months	Low level of chronicity (GCPS); Function (NDI)	R2=34%	Univariate variables: 21+5 category Df Multivariate variables: 21 Predictors in model: 2 Analysed: 128/146 Participants/Predictors: 128/47=2.7
Puentedura et al 2012 [50]; physical therapy USA and Spain outpatient clinics, 2009-2011	Neck pain. n=82, mean age 38,3 years, 59 % female, mean NDI 15,3/50. Cervical spine manipulation and home exercises	Recovery (GROC), second or third session	Symptom duration less than 38 days; Positive expectation that manipulation will help; Side-to-side difference in cervical rotation ROM of 10° or greater; Pain with PA testing of the middle cervical spine.	R2=79%	Univariate variables: 75 Multivariate variables: 9 Predictors in model: 4 Analysed: 82/82 Events/Non-events: 32/50 EPV: 32/84=0.4
Raney et al 2009 [93]; USA army physical therapy centre, 2006-2007	Neck pain-with or without arm pain. n=80; mean age 47.8 (SD10.7), NDI 33.1 (SD 12.7), mean duration 292.4 days. Six standardized physical therapy sessions: cervical traction, exercise, advise to stay active	Recovery (GROC), last visit	Age ≥55; Positive shoulder abduction test; positive ULTT A; Peripheralization on C4-7 PA; positive neck distraction	No information	Univariate variables: >24+ unclear category count Multivariate variables: 15 Predictors in model: 5 Analysed: 68/80 Events/Non-events: 30/38 EPV: 30/>39=<0.8
Rebbeck et al 2006 [94]; Australian injured insurance claimants 2001	WAD; n=250; mean age 39,0; 70% female	Recovery (GPE), 24 months	Initial injury disability score (FRI); Claim status	R2=20%	Univariate variables: unclear Multivariate variables: 12 Predictors in model: 2 Analysed: 56/58 Events/Non-events: 56/58 EPV: 56/12=4.6 (unclear if <4.6)
Ritchie et al 2013 [95]; Australian hospital accident and emergency departments, primary care practices, and general advertisement; 2006-2010	WAD-acute. n=336, mean age 36,4 y; mean VAS pain 4,2. Usual care not withheld from.	Function (NDI), 12 months; model recovery	NDI ≤ 32; Age ≤ 35	R2=16% (Cox & Snell) R2=21% (Nagelkerke)	Univariate variables: 8 Multivariate variables: 8 Predictors in model: 2 Analysed: 262/336 Events/Non-events: 120/142 EPV: 120/16=7.5
		Function (NDI), 12 months; model ongoing disability	NDI ≥ 40; Age ≥ 35; Hyper arousal (PDS subscale ≥ 6)	R2=25% (Cox & Snell) R2=36 % (Nagelkerke)	Univariate variables: 8 Multivariate variables: 8 Predictors in model: 3 Analysed: 262/336 Events/Non-events: 69/193 EPV: 69/16=4.3
Rubinstein et al 2008 [58], Dutch multicentre	Neck pain of any duration. n=529, mean age 41.2 (SD	Pain (NRS), 12 months	Highest level of education; Number of days with neck pain in the preceding year;	No information	Univariate variables: 27+14 category Df Multivariate variables: 25 Predictors in model: 7

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
chiropractic clinics 2004-2005	11.5), 69% female, 75% >12 wk. complaints duration, NDI mean 12.8/50 (SD 6.5). Chiropractic treatment		Intermittent neck pain in the preceding year versus constant; Tiredness; Expected treatment effectiveness.		Analysed: 424/529 Events/Non-events: 284/140 EPV: 140/66=2.1
		Function (NDI), 12 months	Highest level of education; Working status; Fear of -or apprehension concerning- treatment; Headache; Kinesiophobia; Morning pain; No. of days with neck pain in the preceding year; Radiating pain; Tiredness	No information	Univariate variables: 27+14 category Df Multivariate variables: 25 Predictors in model: 11 Analysed: 405/529 Events/Non-events: 134/271 EPV: 134/66=2
		Recovery (GROC), 12 months	Working status; Expected treatment effectiveness; Intermittent neck pain in the preceding year versus constant; Previous episode with neck pain; No. of days with neck pain in the preceding year.	R2=47%	Univariate variables: 27+14 category Df Multivariate variables: 10 Predictors in model: 5 Analysed: 479/529 Events/Non-events: 319/159 EPV: 159/53=3.1
Ssavedra et al 2011 [53]; Spain physical therapy clinic 2009-2010	Neck pain-with or without arm symptoms. n=81; mean age 39.4 (SD 9.2), 70% female; NDI 14.2/50 (SD 5.2). Max 3 manipulation sessions.	Recovery (GROC), second or third session	Pain (NPRS) > 4.5; Extension range of motion < 46°; Hypermobility T1; Negative ULTT; Gender	R2 = 0.38; H-L statistic p=0.38	Univariate variables: 103 Multivariate variables: 8 Predictors in model: 5 Analysed: 81/81 Events/Non-events: 50/31 EPV: 31/111=0.3
Sterling et al 2005 [61]; primary care, emergency centre, advertisement*	WAD acute, grade 2 or 3. n=80; mean age 36.2 (SD12.6) 70% female, mean NDI 34.15 (SD 2.37). Free to pursue any treatment.	Function (NDI), 6 months	Function (NDI); Age; Left rotation range of motion; Cold pain threshold; IES; QI	Adjusted R2=63%	Univariate variables: 21 Multivariate variables: 21 Predictors in model: 6 Analysed: 76/80 Participants/Predictors: 76/21=3.6
		Function (NDI), 6 months; model moderate/severe versus all	Function (NDI); Age, Cold pain threshold; IES	No information	Univariate variables: 21 Multivariate variables: 21 Predictors in model: 4 Analysed: 76/80 Events/Non-events: 17/59 EPV: 17/21=0.8
		Function (NDI), 6 months; model recovered versus mild persistent	Function (NDI); General health (GHQ-28); Extension range of motion	No information	Univariate variables: 21 Multivariate variables: 21 Predictors in model: 6 Analysed: 76/80 Events/Non-events: 29/30 EPV: 29/21=0.9
Sterling et al 2006 [70]; primary care, emergency centre, advertisement	WAD, grade 2 or 3. n=76; mean age 36.27 (SD 12.69) years, 70% female. Free to pursue any treatment	Function (NDI), 2-3 years	Function (NDI); Age; Left rotation range of motion; Cold pain threshold; IES; QI	Adjusted R2=56%	Univariate variables: 22 Multivariate variables: 22 Predictors in model: 7 Analysed: 65/76 Participants/Predictors: 65/22=2.5
		Function (NDI), 2-3 years; model moderate/severe versus all	Function (NDI); Age, Cold pain threshold; IES	No information	Univariate variables: 22 Multivariate variables: 22 Predictors in model: 4 Analysed: 65/76

First author, country, setting, study date	Participants and treatment	Outcomes, follow up	Predictors in final model	Model Performance #	Notes ##
					Events/Non-events: 14/51 EPV: 14/22=0.6
		Function (NDI), 2-3 years; model recovered versus mild persistent	Function (NDI);	No information	Univariate variables: 22 Multivariate variables: 22 Predictors in model: 1 Analysed: 65/76 Events/Non-events: 26/25 EPV: 25/22=1.1
Sterner et al 2003 [45]; Swedish emergency, GP 1997-1998	WAD -acute, grade 1,2,3. n= 356 , mean age 34.1 (SD12.1) years; 52.4% women	Function (NDI), 16 ± 2 months	Previous neck pain; Low educational level; Female gender; WAD grades 2,3.	No information	Univariate variables: 8 Multivariate variables: 8 Predictors in model: 4 Analysed: 296/356 Events/Non-events: 201/95 EPV: 95/8=11.9
Sturzenegger et al 1995 [96], Swiss GP referrals to hospital	WAD-acute. n=137, mean age 30,6 (SD 9,5), 62% woman (analysed group). Usual GP care	Symptoms score, 1 year	Head position; Unpreparedness; Car stationary when hit; Neck pain intensity; Headache intensity	No information	Univariate variables: 30+10 category Df Multivariate variables: 30+10 category Df Predictors in model: 5 Analysed: 117/137 Events/Non-events: 28/89 EPV: 28/40=0.7
Tseng et al 2006 [54]. Taiwan 2 hospitals outpatient dept.	Neck pain. n=100. Mean age 46, 66% female, mean NDI 1,7. Treatment cervical HVT manipulations.	Composite endpoint (NRS or Perceived improvement or Satisfaction level), direct after treatment	Function (NDI <11.50); Bilateral involvement pattern; Not performing sedentary work >5 h/day; Feeling better while moving the neck; Without feeling worse while extending the neck; Diagnosis of spondylosis without radiculopathy	R2=50%	Univariate variables: 30+8 category Df Multivariate variables: 11 Predictors in model: 6 Analysed: 100/100 Events/Non-events: 60/40 EPV: 40/49=0.8
Vos et al 2008 [59]; Dutch GP 2001-2002	Neck pain-acute. n=187, mean age 38.2 (SD13.3), 64% female, mean NDI 14.4/50 (SD 6.5). Standard GP care.	Recovery (7-point ordinal scale), 12 months	Gender; Pain upper neck: Radiating pain to back; Duration >2 wks; GP advised wait and see.	R2=38%	Univariate variables: 19+6 category Df Multivariate variables: 12 Predictors in model: 5 Analysed: 138/187 Events/Non-events: 63/75 EPV: 63/37=1.7
Walton et al 2011 [97]; Canadian outpatient physiotherapy clinics	WAD-acute. n=63 mean age 38.0 (SD14.1) y, 75% female, mean NPRS 5.1 (SD2.4). Standard PT rehabilitation.	Function (NDI), 1-3 months	Gender; Pain (NPRS); Pressure Pain Threshold	R2=38.6%	Univariate variables: 7 Multivariate variables: 3 Predictors in model: 3 Analysed: 45/63 Participants/Predictors: 45/10=4.5
Williamson et al 2015 [98]; UK physiotherapy clinics as sub cohort emergency dept. visitors, 2006-2007	n=599; mean age 39,9 (SD13,1), 63% female, mean NDI 41.8 (SD 16.2)	Function INDI), 12 months	Function (NDI); Predicted time to recovery; Psychological distress (GHQ); Passive coping (CSQ subscale); number of symptoms	R2=19% (Cox and Snell) R2=28% (Nagelkerke)	Univariate variables: 23 Multivariate variables: 9 Predictors in model: 5 Analysed: 430/599 Events/Non-events: 136/223 EPV: 136/32=4.3

* Type 1b internal validation
# For calibration and discrimination with Concordance statistic (c-statistic) or Area Under the Curve (AUC), values in parentheses are 95% Confidence Interval (CI); R2=R-squared statistic; H-L=Hosmer-Lemeshow statistic
Events per variable (EPV) is based on degrees of freedom (Df) used during total modelling process in logistic regression (counts: Df of univariate variables if selected by + Df if categorized + Df in multivariate modelling).

Abbreviations: SD=Standard Deviation; WAD= Whiplash Associated Disorder; IQR=Inter Quartile Range; NASS=North American Spine Society questionnaire; HADS=Hospital Anxiety and Depression Scale; CSQ=Coping Strategies Questionnaire; SF-36/12=Short Form; PDI=Pain Disability Index; GHQ=General Health Questionnaire; NDI=Neck Disability Index; GROC=Global Rating Of Change scale; PTSD=Post Traumatic Stress Disorder; NPRS=Numeric Pain Rating Scale; VAS=Visual Analogue Scale; NPQ=Northwick Park neck pain Questionnaire; FABQ=Fear Avoidance Beliefs Questionnaire; FABQ-PA= Physical Activity subscale; PCS=Pain Catastrophizing Scale; NRS=Numeric Rating Scale; PSFS=Patient Specific Functional Scale; SCL-90=Symptom Checklist; T=Thoracic; ÖMPQ=Örebro Musculoskeletal Pain Questionnaire; NPOS=Neck Pain Outcome Score; SF12 PCS=Pain Catastrophizing subScale; GP=General
Practitioner; GBG=Gargan and Bannister Grade; TSK=Tampa Scale for Kinesiophobia; S-LANSS=Self-reported Leeds Assessment of Neuropathic Symptoms Score; ULNT-median=Upper Limb Neurodynamic Test; BNQ=Bournemouth Neck Questionnaire; IMA=Isometric Muscle Activity; 4DSQ=4 Dimensional Symptom Questionnaire; PCCL=Pain Coping and Control List; PGIC=Patient Global Impression of Change scale; GPE=Global Perceived Effect; GCPS=Graded Chronic Pain Scale; ROM=Range Of Motion; PA=Posterior Anterior; ULTT=Upper Limb Tension Test; C=Cervical; FRI=Functional rating Index; PDS=Posttraumatic Diagnostic Scale; IES=Impact of Events Scale; QI=Quotient of Integrals in blood flow; CPT=Cold Pain Threshold; CSQ=Coping Strategies Questionnaire

Appendix 3. Study characteristics of the validation only and combined studies.

First author, country, setting, study date	Participants and treatment	Model (in validation only studies), Outcomes, follow up	Predictors in final model	Model Performance [a]	Clinical Measures [b]	Notes [b]
Cleland et al 2010 [71]; USA multi-site physiotherapy clinics 2007-2008 [e]	Neck pain. n = 140; mean age 39.9 years (SD11.3), 69% female. Two treatment groups: exercise and exercise plus thoracic manipulation.	Model CPR from Cleland 2007. Pain, Function (NPRS, NDI) over time for treatment group and status on the prediction rule, 1 week, 4 weeks, 6 months	Duration <30 days; No symptoms distal to shoulder; Looking upwards do not aggravate; FABQPA <12; Diminished T3–T5 kyphosis; Cervical extension <30 degrees.	Repeated measures analysis, 3-way interaction for NDI, p = 0.79; for NPRS p = 0.22	Findings does not support prediction rule	Required sample size was not met: analysed: 104/140
Fritz et al 2014 [72]; USA physician and physical therapy offices, 2009-2012 [e]	Neck pain and arm symptoms. n = 86, mean age 46.9 (SD10.7) years, 53,5% female. Three treatment groups: exercise, exercise plus traction, exercise plus home traction.	Model CPR from Raney 2009. Recovery (GROC) over time for traction most effective according to status on the prediction rule, 4 weeks, 6 months, 12 months;	Age ≥55; Positive shoulder abduction test; positive ULTT A; Peripheralization on C4-7 PA; positive neck distraction Positive if ≥3 or more factors, negative if ≤ 2	Repeated measures analysis, 3-way interaction at 6 months, for NDI, p = 0.07; for NPRS p = 0.77	Findings does not support prediction rule	Required sample size was not met
Keating et al 2005 [74]; Australian physiotherapy clinic 1998-2002 [e]	Neck pain-chronic. **Derivation**: n = 336, mean age 41.3 yrs, 65% female, mean symptoms 36 mths, median NDI 34**.** Received supervised exercise program.	Function (NDI), at discharge	Function: NDI lifting; NDI reading	No information	PPV 64% and NPV 74%	Univariate variables: 39+>50 category Df Multivariate variables: 9 Predictors in model: 2 Analysed: 97/122 Events/Non-events: 54/43 EPV: 43/>89 = <0.5
	Neck pain-chronic. **Validation:** n = 214, mean age 40.5 yrs, 67% female, mean symptoms 60 mths, median NDI 36. Received supervised exercise program.			No information	PPV 56% and no improvement of prediction, findings does not support prediction rule. NPV 74% and some improvement of prediction	No information required sample size.
Radanov et al 1996 [51]; Swiss GP **(derivation);** Swiss insurance company **(validation)** [d]	**Derivation:** WAD. n = 117, mean age 30,7 (SD9,6), 58% women	Recovery, 12 months	Impaired neck movement; Pretraumatic headache; History of head trauma; Age; Intensity neck pain; Intensity headache; Nervousness score; Neuroticism score; Focus attention score.	No information	96% correctly predicted in derivation group	Univariate variables: 60+10 category Df Multivariate variables: unclear Predictors in model: 9 Analysed: 117/117 Events/Non-events: 89/28 EPV: 28/>70 = 0.4
	Validation: n = 16, mean age 32,5 (SD8,8), 25% women			No information	88% correctly predicted in validation group	No information required sample size. Sample size was not met with n-16
Ritchie et al 2015 [73]; Australian hospital accident and emergency departments, primary care practices, and general advertisement; 2006-2010 [e]	WAD acute, grade 2. n = 101 **Cohort 1:** N = 53 mean age 33.4 (SD 9.4), NDI 33.8% SD (19.6) **Cohort 2**: n = 48 mean age 35.2 (SD 11.9), NDI 29.0% (SD 17.1).	Model recovery from Ritchie 2013. Function (NDI), 6 months	NDI ≤ 32, Age ≤ 35.	H-L statistic p = 0.81	Sensitivity 54.9 (40.5-68.6), Specificity 86.0 (72.6-93.7), +LR 3.9 (1.9-8.1), -LR 0.5 (0.4-0.7). PVV 80.0 (62.5-91.7). Prevalence 50%	Required sample size was met. Events/Non-events: 51/101 Analysed: 101/101
		Model ongoing disability from Ritchie 2013. Function (NDI), 6 months	NDI ≥ 40 Age ≥ 35, Hyper arousal (PDS subscale ≥ 6)	H-L statistic p = 0.65	Sensitivity 43.5 (22.9-65.1), Specificity 98.7 (92.9-99.9), +LR 33.9 (4.6-251.2), -LR 0.6 (0.4-0.8) PPV 90.9 (58.7-98.5). Prevalence 23%	Required sample size was met. Events/Non-events: 23/101 Analysed: 101/101
Schellingerhout et al 2010 [52], Dutch primary care settings **(derivation);** UK primary care settings **(validation)** [d]	Neck pain-nonspecific. **Derivation:** n = 468 mean age 45.4 (SD 11.8), 61% female, NDI 14.5/50 (SD 6.7). Three combined RCTs: GP, PT, MT, graded activity	Recovery (GROC), 6 months	Age, pain intensity, headache radiation of pain to elbow/shoulder, previous neck complaints, cause of complaints, low back pain, employment status, euroQOL 100 mm VAS.	R2 = 12%; AUC = 0.66 (0.61-0.71), For score chart: R2 = 12%; AUC = 0.66 (0.62-0.71) Calibration plot; H-L statistic p = 0.61.	Accuracy statistics for various scores on chart. Sensitivity, Specificity, PVV e.g. for score ≥35: Sensitivity 0.61 (0.55-0.68), Specificity 0.61 (0.55-0.67), PPV 54% (47%-60%). Prevalence 43% (38%-47%)	Univariate variables: 17+6 category Df Multivariate variables: 10+5 category Df Predictors in model: 9 Analysed: 468/468 Events/Non-events: 199/269 EPV: 199/38 = 5.2

	Validation: n = 315 mean age 48.8 (SD 12.1), 64% female. MT and electrotherapy	Recovery (GROC), 6 months		AUC 0.65 (0.59-0.71). For score chart R2 = 10%; AUC 0.66 (0.59-0.72) Calibration plot; H-L statistic p = 0.61.	Accuracy statistics for various scores on chart. Sensitivity, Specificity, PVV e.g. for score ≥35: Sensitivity 0.63(0.54-0.71), Specificity 0.60 (0.53-0.67), PPV 51% (53%-59%). Prevalence 39% (34%-43%)	No information required sample size.
Sterling et al 2012 [41]; Multisite: Australia, Canada, Iceland; primary care, emergency, advertisement, 2005-2008 [e,f]	WAD acute, grade 1,2,3. Vmodel: n = 286, mean age 35.3 (SD13.08) years, 62,6% female, 29% >30/100 NDI. Free to pursue any treatment.	Model from Sterling 2005. Function (NDI), 12 months	Initial NDI score; Age; ROM Left rotation; Cold pain threshold; IES; QI (blood flow Quotient of Integrals)	R2 = 50%; AUC 0.85 (0.79–0.91) Calibration plot	Accuracy statistics for 2 specific NDI cut points: NDI 32: Sensitivity 80%, Specificity 71%; NDI 29: Sensitivity 90%, Specificity 63%	Required sample size was met.
		Regression model currently developed using same 6 predictors. Function (NDI), 12 months	NDI; Cold pain threshold	R2 = 50%; AUC 0.89 (0.84–0.94)	Accuracy statistics for 2 specific NDI cut points: Sensitivity, Specificity, PVV, NPV	Multivariate variables: 6 Predictors in model: 2 Analysed: 225/286 Participants/Predictors: 225/2 = 113
		Regression model currently developed adjusted for site. Function (NDI), 12 months	NDI; Age; Cold pain threshold, IES	R2 = 56%; AUC 0.91 (0.86–0.95)	Accuracy statistics for 2 specific NDI cut points: Sensitivity, Specificity, PVV, NPV	Multivariate variables: 7 Predictors in model: 4 Analysed: 225/286 Participants/Predictors: 225/4 = 286

[a] For calibration and discrimination with Cocordance statistic (c-statistic) or Area Under the Curve (AUC), values in parentheses are 95% Confidence Interval (CI), R^2 = R-squared statistic, and H-L = Hosmer-Lemeshow statistic.

[b] Events per variable (EPV) is based on degrees of freedom (Df) used during total modelling process in logistic regression (counts: Df of univariate variables if selected by + Df if categorized + Df in multivariate modelling).

[c] Type 2b study, intermediate (temporal) validation

[d] Type 3 study, development and validation using separate data set

[e] Type 4 study, validation only

[f] Contains 2 regression models developed in validation study

Abbreviations: C = Cervical; CPR = Clinical Prediction Rule; CPT = Cold Pain Threshold; EuroQOL = Quality Of Life; FABQ = Fear Avoidance Beliefs Questionnaire; FABQ-PA = Physical Activity subscale; GP = General Practitioner; GROC = Global Rating Of Change scale; IES = Impact of Events Scale; MT = Manual Therapy; NDI = Neck Disability Index; NPRS = Numeric Pain Rating Scale; NPV = Negative Predictive Value; NRS = Numeric Rating Scale; PA = Posterior Anterior; PDS = Posttraumatic Stress Diagnostic Scale; PPV = Positive Predictive Value; PSFS = Patient Specific Functional Scale; PT = Physical Therapy; QI = Quotient of Integrals in blood flow; ROM = Range Of Motion; T = Thoracic; ULTT = Upper Limb Tension Test; VAS = Visual Analogue Scale; WAD = Whiplash Associated Disorder; +LR = Positive Likelihood Ratio; -LR = Negative Likelihood Ratio.

References

1. Hoy D, March L, Woolf A, Blyth F, Brooks P, Smith E, et al. The global burden of neck pain: Estimates from the global burden of disease 2010 study. Ann Rheum Dis. 2014;73:1309–1315.
2. Vos T, Allen C, Arora M, Bhutta Z, Brown A, Carter A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388:1545–1602.
3. 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;92:824–829.
4. 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.
5. van der Velde G, Yu H, Paulden M, Côté P, Varatharajan S, Shearer HM, et al. Which interventions are cost-effective for the management of whiplash-associated and neck pain-associated disorders? A systematic review of the health economic literature by the Ontario Protocol for Traffic Injury Management (OPTIMa) Collab-oration. Spine J. 2016;16:1582–1597.
6. Vincent K, Maigne J-Y, Fischhoff C, Lanlo O, Dagenais S. Systematic review of manual therapies for nonspecific neck pain. Joint Bone Spine. 2013;80:508–515.
7. Gross A, Kay T, Paquin J, Blanchette S, Lalonde P, Christie T, et al. Exercises for mechanical neck disorders (Review). Cochrane Database Syst Rev. 2015;(1).
8. Hurwitz EL, Carragee EJ, van der Velde G, Carroll LJ, Nordin M, Guzman J, et al. Treatment of Neck Pain: Noninvasive Interventions. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. J Manipulative Physiol Ther. 2009;32:S141–S175.
9. Fritz JM, Brennan GP. Preliminary examination of a proposed treatment-based classification system for patients receiving physical therapy interventions for neck pain. Phys Ther. 2007;87:513–524.
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. Eur Spine J. 2008;17:75–82.
11. Cleland JA, Childs JD, Fritz JM, Whitman JM, Eberhart SL. Development of a clinical prediction rule for guiding treatment of a subgroup of patients with neck pain: use of thoracic spine manipulation, exercise, and patient education. Phys Ther. 2007;87:9–23.
12. Carroll LJ, Hurwitz EL, Côté P, Hogg-Johnson S, Carragee EJ, Nordin M, et al. Research priorities and methodological implications: the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. J Manipulative Physiol Ther. 2009;32:244–251. 13. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ. 2013;346:e5595–e5595.
14. Moons KGM, Biesheuvel CJ, Grobbee DE. Test Research versus Diagnostic Research. Clin Chem. 2004;50:473–476.
15. Moons KGM, Altman DG, Reitsma JB, Collins GS. New guideline for the reporting of studies developing, validating, or updating a prediction model. Clin Chem. 2015;61:565–566.
16. van Trijffel E, Lindeboom R, Bossuyt PM, Schmitt MA, Lucas C, Koes BW, et al. Indicating spinal joint mobilisations or manipulations in patients with neck or low- back pain: protocol of an inter-examiner reliability study among manual therapists. Chiropr Man Therap. 2014;22:22.
17. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375–b375.
18. 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.
19. Steyerberg E, Moons KGM, van der Windt D, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) series 3: prognostic models. PLoS Med. 2013;10:e1001381.
20. van Oort L, van den Berg T, Koes BW, de Vet RHCW, Anema HJR, Meymans MW, et al. Preliminary state of development of prediction models for primary care physical therapy: A systematic review. J Clin Epidemiol. 2012;65:1257–1266.
21. Stanton TR. Critical appraisal of clinical prediction rules that aim to optimize treat- ment selection for musculoskeletal conditions. Am Phys Ther Assoc. 2010;90:177–201.
22. Beneciuk JM, Bishop MD, George SZ. Clinical prediction rules for physical therapy interventions: a systematic review. Am Phys Ther Assoc. 2009;89:114–124.
23. Kelly J, Ritchie C, Sterling M. Clinical prediction rules for prognosis and treatment prescription in neck pain: A systematic review. Musculoskelet Sci Pract. 2017;27:155–164.
24. Ensor J, Riley RD, Moore D, Snell KIE, Bayliss S, Fitzmaurice D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. BMJ Open. 2016;6:e011190.
25. Braun C, Hanchard NC, Batterham AM, Handoll HH, Betthauser A. Prognostic models in adults undergoing physical therapy for rotator cuff disorders: systematic review. Am Phys Ther Assoc. 2016;96:961–971.
26. Halligan S, Boone D, Bhatnagar G, Ahmad T, Bloom S, Rodriguez-Justo M, et al. Prognostic biomarkers to identify patients destined to develop severe Crohn's disease who may benefit from early biological therapy: protocol for a systematic review, meta-analysis and external validation. Syst Rev. 2016;5:206.
27. Ingui BJ, Rogers MAM. Searching for clinical prediction rules in MEDLINE. J Am Med Informatics Assoc. 2001;8:391–397.
28. Pillastrini P, Vanti C, Curti S, Mattioli S, Ferrari S, Violante FS, et al. Using PubMed search strings for efficient retrieval of manual therapy research literature. J Manipulative Physiol Ther. 2015;38:159–166.
29. Kwon Y, Lemieux M, McTavish J, Wathen N. Identifying and removing duplicate records from systematic review searches. J Med Libr Assoc. 2015;103:184–188.
30. Clark GM. Prognostic factors versus predictive factors: Examples from a clinical trial of erlotinib. Mol Oncol. 2008;1:406–412.
31. Hidalgo B, Goodman M. Multivariate or multivariable regression? Am J Public Health. 2013;103:39–40.
32. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer Science and Business Media; 2009.
33. Guzman J, Hurwitz EL, Carroll LJ, Haldeman S, Côté P, Carragee EJ, et al. A New Conceptual Model of Neck Pain. Eur Spine J. 2008;17:14–23.
34. 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. 2008;32:S14–S23.
35. Hoving JL, de Vet HCW, Koes BW, van der Windt DA, Assendelft WJ, van Mameren H, et al. Manual therapy, physical therapy, or continued care by the general practi-tioner for patients with neck pain: long-term results from a pragmatic randomized clinical trial. Clin J Pain. 2006;22:370–377.

36. Wolff R, Whiting P, Mallet S, Riley R, Westwood M, Kleijnen K, et al. PROBAST: a risk of bias tool for prediction modelling studies | The 23rd Cochrane Colloquium. http://2015.colloquium.cochrane.org/abstracts/probast-risk-bias-tool-prediction- modelling-studies. Published 2015 [accessed 28/07/2016].
37. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res. 2017;26:796–808.
38. 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:e1001744. 39. Moons KGM, 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: W1–W73.
40. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explana- tory variable. BMC Med Res Methodol. 2012;12:82.
41. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken, NJ: Wiley; 2013.
42. Sterling M, Hendrikz J, Kenardy J, Kristjansson E, Dumas JP, Niere K, et al. Assess-ment and validation of prognostic models for poor functional recovery 12 months after whiplash injury: A multicentre inception cohort study. Pain. 2012;153: 1727–1734. 43. Cobo EP, Mesquida MEP, Fanegas EP, Atanasio EM, Pastor MBS, Pont CP, et al. What factors have influence on persistence of neck pain after a whiplash? Spine. 2010;35: E338–E343.
44. Kyhlbäck M, Thierfelder T, Söderlund A. Prognostic factors in whiplash-associated disorders. Int J Rehabil Res. 2002;187:181–187.
45. Lankester BJA, Garneti N, Gargan MF, Bannister GC. Factors predicting outcome after whiplash injury in subjects pursuing litigation. Eur Spine J. 2006;15: 902–907.
46. Sterner Y, Toolanen G, Gerdle B, Hildingsson C. The incidence of whiplash trauma and the effects of different factors on recovery. J Spinal Disord. 2003;16:195–199.
47. Cai C, Ming G, Ng LY. Development of a clinical prediction rule to identify patients with neck pain who are likely to benefit from home-based mechanical cervical traction. Eur Spine J. 2011;20:912–922.
48. Hanney WJ, Kolber MJ, George SZ, Young I, Patel CK, Cleland JA. Development of a preliminary clinical prediction rule to identify patients with neck pain that may benefit from a standardized program of stretching and muscle perfor- mance exercise: a prospective cohort study. Int J Sports Phys Ther. 2013;8: 756–776.
49. Hill JC, Lewis M, Sim J, Hay EM, Dziedzic K. Predictors of poor outcome in patients with neck pain treated by physical therapy. Clin J Pain. 2007;23:683–690.
50. Landers MR, Creger RV, Baker CV, Stutelberg KS, Landers M, Creger R, et al. The use of fear-avoidance beliefs and non organic signs in predicting prolonged disability in patients with neck pain. Man Ther. 2008;13:239–248.
51. Puentedura EJ, Cleland JA, Landers MR, Mintken P, Louw A, Fernandez-de-Las-Penas C. Development of a clinical prediction rule to identify patients with neck pain likely to benefit from thrust joint manipulation to the cervical spine. J Orthop Sports Phys Ther. 2012;42:577–592.
52. Radanov BP, Sturzenegger M. Predicting recovery from common whiplash. Eur Neurol. 1996;36:48–51.
53. Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HCW, Koes BW. Prognosis of patients with nonspecific neck pain: development and external validation of a prediction rule for persistence of complaints. Spine. 2010;35: E827–E835.

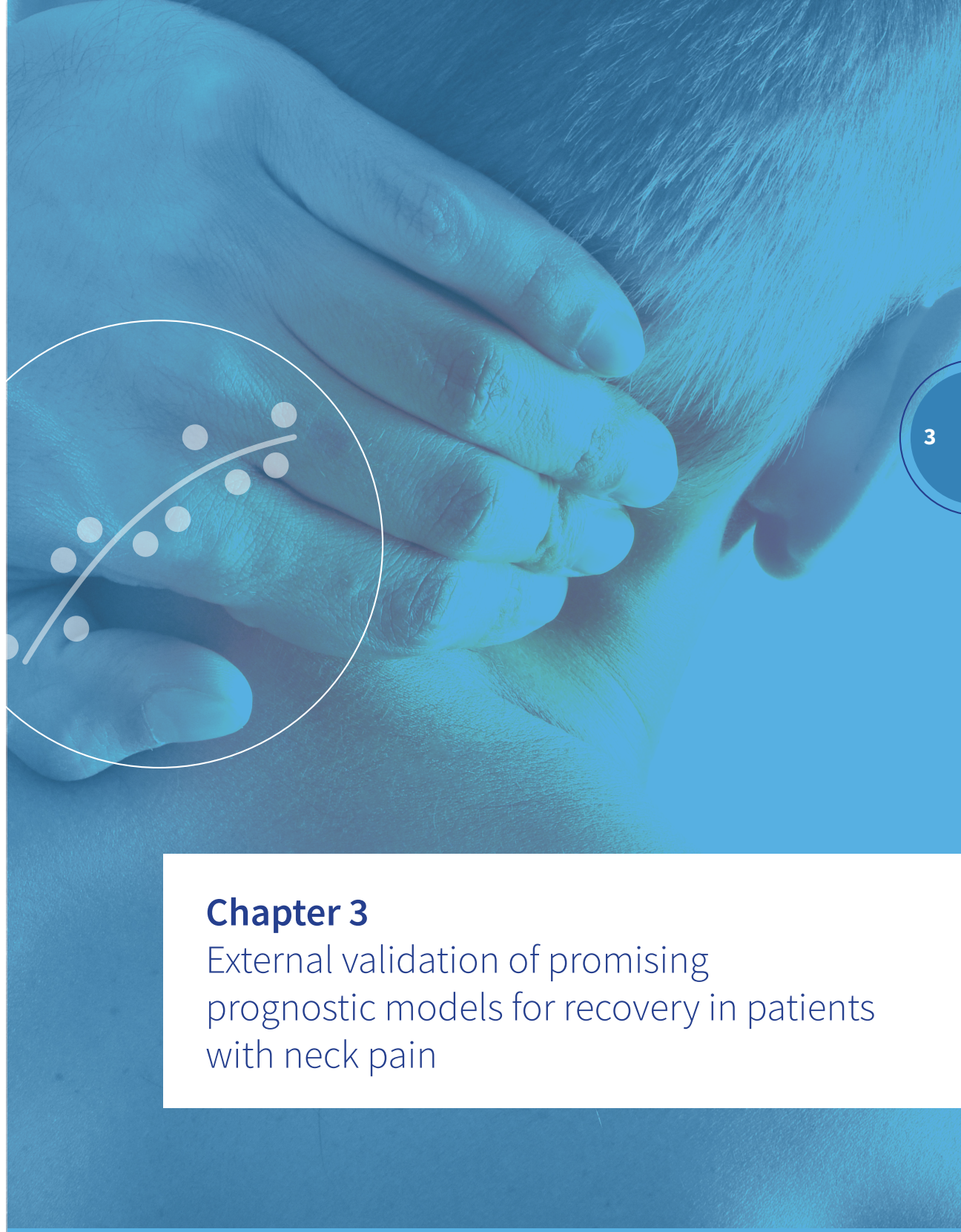
54. Saavedra-Hernández M, Castro-Sánchez AM, Fernández-De-Las-Peñas C, Cleland JA, Ortega-Santiago R, Arroyo-Morales M. Predictors for identifying patients with mechanical neck pain who are likely to achieve short-term success with manipu-lative interventions directed at the cervical and thoracic spine. J Manipulative Physiol Ther. 2011;34:144–152.
55. Tseng Y, Wang WTJ, Chen W-Y, Hou T-J, Chen T-C, Lieu F-K. Predictors for the immediate responders to cervical manipulation in patients with neck pain. Man Ther. 2006;11:306–315.
56. Angst F, Gantenbein AR, Lehmann S, Gysi-Klaus F, Aeschlimann A, Michel BA, et al. Multidimensional associative factors for improvement in pain, function, and working capacity after rehabilitation of whiplash associated disorder: a prognostic, prospective outcome study. BMC Musculoskelet Disord. 2014;15:130.
57. Chiarotto A, Fortunato S, Falla D. Predictors of outcome following a short multi-modal rehabilitation program for patients with whiplash associated disorders. Eur J Phys Rehabil Med. 2015;51:133–141.
58. Hoving JL, de Vet HCW, Twisk JWR, Deville WLJM, van der Windt DAWM, Koes BW. Prognostic factors for neck pain in general practice. Pain. 2004;110:639–645.
59. Rubinstein SM, Knol DL, Leboeuf-Yde C, de Koekkoek TE, Pfeifle CE, van Tulder MW. Predictors of a favorable outcome in patients treated by chiropractors for neck pain. Spine. 2008;33:1451–1458.
60. Vos CJ, Verhagen AP, Passchier J, Koes BW. Clinical course and prognostic factors in acute neck pain: an inception cohort study in general practice. Pain Med. 2008;9:572–580.
61. 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.
62. Sterling M, Jull G, Vicenzino B, Kenardy J, Darnell R. Physical and psychological factors predict outcome following whiplash injury. Pain. 2005;114:141–148. 63. Hendriks EJM, Scholten-Peeters GGM, Van Der Windt DAWM, Neeleman-Van Der Steen CWM, Oostendorp RAB, Verhagen AP. Prognostic factors for poor recovery in acute whiplash patients. Pain. 2005;114:408–416. 64. Carstensen TBW, Fink P, Oernboel E, Kasch H, Jensen TS, Frostholm L. Sick leave within 5 years of whiplash trauma predicts recovery: a prospective cohort and register-based study. PLoS One. 2015;10:e0130298. 65. Nederhand MJ, Ijzerman MJ, Hermens HJ, Turk DC, Zilvold G. Predictive value of fear avoidance in developing chronic neck pain disability: Consequences for clinical decision making. Arch Phys Med Rehabil. 2004;85:496–501.
66. Nee RJ, Vicenzino B, Jull GA, Cleland JA, Coppieters MW. Baseline characteristics of patients with nerve-related neck and arm pain predict the likely response to neural tissue management. J Orthop Sports Phys Ther. 2013;43:379–391.
67. Peterson C, Bolton J, Humphreys BK. Predictors of outcome in neck pain patients undergoing chiropractic care: comparison of acute and chronic patients. Chiropr Man Therap. 2012;20:27.
68. Åsenlöf P, Bring A, Söderlund A. The clinical course over the first year of Whiplash Associated Disorders (WAD): Pain-related disability predicts outcome in a mildly affected sample. BMC Musculoskelet Disord. 2013;14.
69. Baltov P, Cote J, Truchon M, Feldman DE. Psychosocial and socio-demographic factors associated with outcomes for patients undergoing rehabilitation for chronic whiplash associated disorders: a pilot study. Disabil Rehabil. 2008;30:1947–1955.
70. Dagfinrud H, Storheim K, Magnussen LH, Ødegaard T, Hoftaniska I, Larsen LG, et al. The predictive validity of the Orebro Musculoskeletal Pain Questionnaire and the clinicians' prognostic assessment following manual therapy treatment of patients with LBP and neck pain. Man Ther. 2013;18:124–129.

71. Sterling M, Jull G, Kenardy J. Physical and psychological factors maintain long-term predictive capacity post-whiplash injury. Pain. 2006;122:102–108.
72. Cleland JA, Mintken PE, Carpenter K, Fritz JM, Glynn P, Whitman J, et al. Examination of a clinical prediction rule to identify patients with neck pain likely to benefit from thoracic spine thrust manipulation and a general cervical range of motion exercise: multi-center randomized clinical trial. Phys Ther. 2010;90:1239–1250.
73. Fritz JM, Thackeray A, Brennan GP, Childs JD. Exercise only, exercise with mechanical traction, or exercise with over-door traction for patients with cervical radiculo-pathy, with or without consideration of status on a previously described subgrouping rule: a randomized clinical trial. J Orthop Sports Phys Ther. 2014;44:45–57.
74. Ritchie C, Hendrikz J, Jull G, Elliott J, Sterling M. External validation of a clinical prediction rule to predict full recovery and ongoing moderate/severe disability following acute whiplash injury. J Orthop Sports Phys Ther. 2015;45:242–250.
75. Keating JL, Kent P, Davidson M, Duke R, McKinnon L, De Nardis R. Predicting short-term response and non-response to neck strengthening exercise for chronic neck pain. J Whiplash Relat Disord. 2005;4:43–55.
76. Haskins R, Osmotherly PG, Rivett DA. Validation and impact analysis of prognostic clinical prediction rules for low back pain is needed: a systematic review. J Clin Epidemiol. 2015;68:821–832.
77. Patel S, Friede T, Froud R, Evans DW, Underwood M. Systematic review of randomized controlled trials of clinical prediction rules for physical therapy in low back pain. Spine. 2013;38:762–769.
78. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66:411–421.
79. Gross AR, Paquin JP, Dupont G, Blanchette S, Lalonde P, Cristie T, et al. Exercises for mechanical neck disorders: A Cochrane review update. Man Ther. 2016;24:25–45.
80. Sutton D, Cote P, Wong JJ, Varatharajan S, Randhawa K, Yu H, et al. Is multimodal care effective for the management of patients with whiplash-associated disorders or neck pain and associated disorders? A systematic review by the Ontario Protocol for Traffic Injury Management (OPTIMa) Collaboration. Spine J. 2014;34–61.
81. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691–698.
82. Atherton K, Wiles NJ, Lecky FE, Hawes SJ, Silman AJ, Macfarlane GJ, et al. Predictors of persistent neck pain after whiplash injury. Emerg Med J. 2006;23:195–201.
83. Buitenhuis J, de Jong PJ, Jaspers JPC, Groothoff JW. Relationship between posttrau-matic stress disorder symptoms and the course of whiplash complaints. J Psychosom Res. 2006;61:681–689.
84. Bunketorp L, Lindh M, Carlsson J, Stener-Victorin E. The perception of pain and pain-related cognitions in subacute whiplash-associated disorders: Its influence on prolonged disability. Disabil Rehabil. 2006;28:271–279.
85. Cecchi F, Molino-Lova R, Paperini A, Boni R, Castagnoli C, Gentile J, et al. Predictors of short- and long-term outcome in patients with chronic non-specific neck pain undergoing an exercise-based rehabilitation program: a prospective cohort study with 1-year follow-up. Intern Emerg Med. 2011;6:413–421.
86. Cleland JA, Fritz JM, Whitman JM, Heath R. Predictors of short-term outcome in people with a clinical diagnosis of cervical radiculopathy. Phys Ther. 2007;87: 1619–1632.
87. Gun RT, Osti OL, O'Riordan A, Mpelasoka F, Eckerwall CGM, Smyth JF. Risk factors for prolonged disability after whiplash injury: a prospective study. Spine. 2005;30:386–391.

88. Hartling L, Pickett W, Brison RJ. Derivation of a clinical decision rule for whiplash associated disorders among individuals involved in rear-end collisions. Accid Anal Prev. 2002;34:531–539.
89. Kyhllman G, Skargren E, Oberg B. Prognostic factors for perceived pain and function at one-year follow-up in primary care patients with neck pain. Disabil Rehabil. 2002;24:364–370.
90. Michaelson P, Sjolander P, Johansson H. Factors predicting pain reduction in chronic back and neck pain after multimodal treatment. Clin J Pain. 2004;20:447–454.
91. Nieto R, Miro J, Huguet A. Pain-related fear of movement and catastrophizing in whiplash-associated disorders. Rehabil Psychol. 2013;58:361–368.
92. Pape E, Brox JI, Hagen KB, Natvig B, Schirmer H. Prognostic factors for chronic neck pain in persons with minor or moderate injuries in traffic accidents. Accid Anal Prev. 2007;39:135–146.
93. Pool JJM, Ostelo RWJG, Knol D, Bouter LM, de Vet HCW. Are psychological factors prognostic indicators of outcome in patients with sub-acute neck pain? Man Ther. 2010;15:111–116.
94. Raney NH, Petersen EJ, Smith TA, Cowan JE, Rendeiro DG, Deyle GD, et al. Development of a clinical prediction rule to identify patients with neck pain likely to benefit from cervical traction and exercise. Eur Spine J. 2009;18:382–391.
95. Rebbeck T, Sindhusake D, Cameron ID, Rubin G, Feyer AM, Walsh J, et al. A prospective cohort study of health outcomes following whiplash associated dis-orders in an Australian population. Inj Prev. 2006;12:93–98.
96. 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;154:2198–2206.
97. Sturzenegger M, Radanov BP, Di Stefano G. The effect of accident mechanisms and initial findings on the long-term course of whiplash injury. J Neurol. 1995;242: 443–449.
98. Walton DM, Macdermid JC, Nielson W, Teasell RW, Reese H, Levesque L. Pressure pain threshold testing demonstrates predictive ability in people with acute whip-lash. J Orthop Sports Phys Ther. 2011;41:658–665.
99. Williamson E, Williams MA, Gates S, Lamb SE. Risk factors for chronic disability in a cohort of patients with acute whiplash associated disorders seeking physiotherapy treatment for persisting symptoms. Physiotherapy. 2015;101:34–43.

Websites

PROBAST www.systematic-reviews.com/probast



Chapter 3
External validation of promising prognostic models for recovery in patients with neck pain

Chapter 3. External validation of promising prognostic models for recovery in patients with neck pain

Roel W. Wingbermühle, Martijn W. Heymans, Emiel van Trijffel, Alessandro Chiarotto, Bart Koes, Arianne P. Verhagen

Brazilian Journal of Physiotherapy. 2021 Nov; 25 (6): 775-784

Abstract

Background: Neck pain is one of the leading causes of disability in most countries and it is likely to increase further. Numerous prognostic models for people with neck pain have been developed, but few have been validated. In a recent systematic review, external validation of three promising models was advised before they can be used in clinical practice. **Objective:** The purpose of this study was to externally validate three promising models that predict neck pain recovery in primary care. **Methods:** This validation cohort consisted of 1311 patients with neck pain of any duration who were prospectively recruited and treated by 345 manual therapists in the Netherlands. Outcome measures were disability (Neck Disability Index) and recovery (Global Perceived Effect Scale) post-treatment and at 1-year follow-up. The assessed models were an Australian Whiplash-Associated Disorders (WAD) model (Amodel), a multicenter WAD model (Mmodel), and a Dutch nonspecific neck pain model (Dmodel). Models' discrimination and calibration were evaluated. **Results:** The Dmodel and Amodel discriminative performance (AUC < 0.70) and calibration measures (slope largely different from 1) were poor. The Mmodel could not be evaluated since several variables nor their proxies were available. **Conclusions:** External validation of promising prognostic models for neck pain recovery was not successful and their clinical use cannot be recommended. We advise clinicians to underpin their current clinical reasoning process with evidence-based individual prognostic factors for recovery. Further research on finding new prognostic factors and developing and validating models with up-to-date methodology is needed for recovery in patients with neck pain in primary care.

Introduction

Neck pain is common and one of the leading causes of disability in most countries.[1,2] From 2005 to 2015, the prevalence of chronic neck pain has increased globally by 21.1% and is likely to increase further.[1,2] Recovery from neck pain related disability mainly takes place in the first few weeks without further subsequent improvement.[3] Acute neck pain prognosis may be even worse than currently recognized which underlines the importance of neck pain prognosis at intake in primary care.[3]
Short-term beneficial effects and cost-effectiveness of non-invasive primary care treatment have been reported but long-term effects are still limited.[4-7] Prognostic models are obtained by multivariable regression and aim to improve the quality of care

for *individual* patients by estimating the probability of a future health outcome or condition being present by combining *patient-specific values* of multiple predictors.[8] Accurate prognostic models can be useful for clinicians to support clinical decisions and for research to risk-stratify participants for clinical trials.[8-10] Compared to derivation studies, models usually perform less well in external validation studies and it is recommended first to test models' generalizability and transportability to evaluate whether their predictive performance remains accurate before broad clinical use can be advised.[11-13]
Numerous prognostic models for people with neck pain have been developed, however, few have been validated.[14-16] In a recent systematic review, three promising models that predict recovery of people with neck pain in primary care were identified.[17] However, their broad clinical use could not be recommended and further external validation was advised.[17] Therefore, the research question of this study was: can these three models be externally validated in a cohort of people with nonspecific neck pain treated with manual therapy in Dutch primary care?

Methods

This external validation study including its statistical analysis was performed according to an a priori constructed and approved study protocol complying with internal university procedures. The included models were: 1) the Australian two-way model (Amodel)[18] predicting full recovery and ongoing moderate to severe disability, measured with the Neck Disability Index (NDI) in patients with Whiplash-Associated Disorders (WAD); 2) the multicenter model (Mmodel)[19] also predicting disability measured with the NDI in patients with WAD, and 3) the Dutch model (Dmodel)[20] predicting recovery measured with a Global Perceived Effect Scale (GPES) in patients with non-specific neck pain. Models' characteristics are presented in Table 1. The findings of this study were reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) recommendations.[21]

ANIMO validation cohort

For validation, existing data from the 'Amersfoorts Nekonderzoek of the Master manuele therapie Opleiding' (ANIMO) study was used. Ethics approval was obtained from Erasmus Medical Centre, Rotterdam, the Netherlands (MEC-2007- 359). The dataset used and analyzed during the current study is available upon reasonable request. ANIMO is a prospective cohort study that aimed to describe usual care manual therapy for patients with neck pain in the Netherlands and explored outcomes and adverse events of treatment. Patients between 18 and 80 years with neck pain consulting a directly accessible manual therapist were recruited from October 2007 until March 2008. Participants with signed informed consent and treatment indication who submitted baseline data were eligible for participation (n = 1193). Received treatment consisted of usual care manual therapy and may have included specific joint mobilizations, high-velocity thrust techniques, myofascial techniques, giving advice, or specific exercises. Further study characteristics are described in detail elsewhere.[22]

Measurement procedure

Participants completed socio-demographic characteristics and questionnaires at baseline, immediately post-treatment, and at 12 months. Manual therapists were blinded from information gathered by patients' questionnaires. At baseline, patients' age, sex, marital status, employment, neck pain duration, neck pain localization, earlier episodes, associated symptoms, current medication, current smoking, current sport, imaging results, additional diagnostics, medical diagnosis, and comorbidities were recorded. Disability was measured using the Dutch versions of the NDI (scale 0-50)[23,24] and the Neck Bournemouth Questionnaire (NBQ, scale 0-70)[25]; pain intensity was measured with a 10-point Numeric Rating Scale (NRS, scale 1-10), and pain-related fear was measured with the Dutch version of the Fear Avoidance Beliefs Questionnaire (FABQ-DV, scale 0-96).[26] Outcomes were measured post-treatment at discharge (mean treatment duration 37.9 days, mean number of 4.3 sessions) and at 12 months follow-up, using the NDI and a GPES (7- point Likert scale).

Validation procedure

Based on models' predictors available in ANIMO, the Amodel(s) and Dmodel were suitable for validation.[20,27] The Mmodel was considered not suitable due to four variables not collected in ANIMO (i.e. cold pain threshold, the impact of events scale, the quotient of a sympathetic vasoconstrictor response; left rotation) with lack of appropriate proxy measures.[28] As the Amodel(s) were developed for people with WAD and ANIMO also contained patients with non-traumatic neck pain, we created a subset of patients with self-reported trauma in ANIMO. We used the NBQ anxious subscale with comparable cutoff value as a proxy for the hyperarousal subscale of the Posttraumatic stress Diagnostic Scale (PDS) because the PDS was not available in ANIMO. For the Dmodel, we removed the quality of life variable (EuroQoL, beta value 0.005) because this was not available in ANIMO.

Table 1 Models' characteristics.

	First author and year	Setting	Condition, treatment and number of participants	Participants characteristics	Outcomes, follow up	Models with intercept, predictors and their weights
Amodel	Ritchie et al. 2013	Australian hospital accident and emergency departments, primary care practices, and recruitment from advertisement	WAD-acute, grade 1,2 or 3; usual care not withheld from; *n* = 336	Mean age 36.4 years. Mean VAS pain: 4,2	**Full recovery:** Function at 12 months NDI score multiplied by two and cutoff ≤ 10%	−1.667; 1.856 NDI initial ≤ 32, 0.717 Age ≤ 35
					Ongoing disability: Function at 12 months NDI score multiplied by two and cutoff ≥ 30%	−2.859; 2.013 NDI initial ≥ 40; 0.811Age ≥ 35, 0.796 Hyper arousal subscale (PDS) ≥ 6
Mmodel	Sterling et al. 2005	Australian hospital accident and emergency departments, primary care practices, and recruitment from advertisement	WAD acute, grade 2 or 3; Free to pursue any treatment; *n* = 80	Mean age 36.2 (SD12.6) years. 70% female Mean NDI 34.15 (SD 2.37)	**Persistent neck complaints:** Function at 6 months, NDI score	11.74; 0.387 Initial NDI score; 0.387 Age, −0.178 ROM Left rotation; 0.505 CPT; 0.338 IES; −0.0147 QI
Dmodel	Schellingerhout et al. 2010	Dutch primary care settings	Neck pain nonspecific; different therapy in RCT (usual care GP, PT, MT, graded activity); *n* = 468	Mean age 45.4 (SD 11.8) years. 61% female NDI 14.5/50 (SD 6.7)	**Recovery**: GPRS at 6 months, dichotomized into recovered or much improved and persistent complaints	−1.704; 0.029 Age, −0.042 pain intensity, 0.198 headache, −0.564 radiation of pain to elbow/shoulder, 0.515 previous neck complaints, 0.234 cause of complaints, 0.829 low back pain, 0.372 employment status, 0.005 EuroQoL, 0.116 accompanying headache * pain intensity, −0.376 accompanying headache * previous neck complaints, 0.392 accompanying headache * radiation of pain, −0.815 accompanying headache * employment status

Abbreviations: WAD= Whiplash Associated Disorder; GP=General Practitioner; PT=Physical Therapy; MT=Manual Therapy; NPRS=Numeric Pain Rating Scale; VAS=Visual Analogue Scale; NDI=-Neck Disability Index; GPRS=Global Perceived Recovery scale; EuroQoL=Quality of Life; ROM=Range Of Motion; IES=Impact of Events Scale; QI=Quotient of Intergrals in blood flow; CPT=Cold Pain Threshold. * indicates interaction terms in the regression models.

We used the same outcome cut-off values as the original studies. We examined baseline demographics, models' predictors, and outcome distribution between the models' development studies and ANIMO as means with standard deviations or frequencies or percentages to compare case mix between studies.

Handling of missing values

The ANIMO data contained missing values and we planned to perform several missing value analyses to decide on multiple imputation for main analyses and complete cases for sensitivity analysis. [2932]-

Statistical analysis

Statistical validation of models' performance

We compared observed outcomes to those predicted by the models and analyzed the full original models in ANIMO and based models' performance on discrimination and calibration measures.[10,13,33] The Amodel was analyzed in both the ANIMO trauma subset as well as the whole dataset. We calculated model's linear predictor and individual probability (p (y = 1) =1/ (1 + $e^{-\text{linear predictor}}$)) for all participants immediately post-treatment and at 1 year follow-up.[34]

Discriminative performance

Discriminative performance indicates whether a model is able to distinguish between patients with and without recovery. It is calculated as the concordance (c) statistic which is comparable to the area under the curve (AUC) of the Receiver Operating Characteristic curve (ROC) for binary data.[13,35] We a priori considered discriminative performance acceptable if AUC was ≥ 0.70.[36]

Calibration performance

Calibration performance refers to the agreement between a model's predicted risks and observed event rates.[37] Preferably, this is reflected by calibration-in-the large, a calibration slope, and a calibration plot.[13,38] The Hosmer-Lemeshow goodness of fit test is often performed in validation studies and if the test is not-significant, it should indicate that the model fits the data well.[36] The models were re-estimated in ANIMO on a logit scale with the linear predictor as only predictor to calculate calibration-in-the-large and the calibration slope.[10,13,30] We evaluated calibration as a percentage of deviation from the ideal calibration slope of 1 and the intercept of 0. Calibration plots' probabilities were calculated to allow observation if all decile groups closely fit the ideal 45° line of identity.[10,13] We performed statistical validation procedures using IBM SPSS 24.0 and R (version 3.4.3).

Finally, we checked the number of events in ANIMO for a minimum of 100, as advised for validation studies that predict binary outcomes.[39,40]

Results

Study characteristics

The baseline characteristics from the ANIMO study and the original studies are presented in Table 2.

Amodels

The ANIMO subset consisted of people with any trauma and neck pain duration, whereas the original Amodel study included people with acute neck pain due to a motor vehicle crash only. People in ANIMO were recruited and treated in primary care with manual therapy and people in the original study were allowed to pursue any treatment and were recruited from general advertisement and emergency departments. On average, people in the original study were 4.8 years younger compared to the ANIMO trauma subset, had 17 NDI points higher disability (0-50 scale), and had 0.9 points more pain (0-10 scale).

Dmodel

There were 8.1% fewer male participants in ANIMO compared to the Dmodel derivation study. Duration of the current episode in the Dmodel derivation cohort resulted in 26% more patients categorized as acute and 13.5% more patients categorized as chronic compared to ANIMO. In ANIMO, the average disability at inception was 1.5 NDI points lower and the average neck pain was 2.4 points less on an 11-point Likert scale. For the other variables, there were 8.8% fewer people with headaches and 20.1% fewer with radiating arm pain. In ANIMO, 2.9% more people had a previous neck pain episode, 24.1% more had concomitant low back pain, and 6.1% more people were employed.

Missing data

There were more than 5% missing data for several baseline variables and all outcome measures (Table 2). Little's Missing Completely at Random (MCAR) test was significant at the $p<0.05$ level so we assumed data were not MCAR. Significant differences in means existed for 24 of 91 variables and differences were small indicating Missing at Random (MAR). Explained variation of missingness varied from 11 to 100% and missing variables were to some extent associated with the other ANIMO variables. Therefore, we assumed data were MAR. We applied multiple regression imputation for missing data using all possible predictors and outcomes, as computationally feasible.[29,31,41] We used the Multivariate Imputation by Chained Equations (MICE) procedure and generated 20 imputed sets.[42]
Regression coefficient estimates and standard errors were pooled using Rubin's Rules and validation performance measures were estimated in each of the 20 completed datasets and then combined using the median.[30,43] We used imputed data for main analyses and complete cases for sensitivity analysis.

Table 2 The baseline characteristics of participants in the ANIMO validation cohort and the original studies.

	ANIMO Validation cohort (*n* = 1193)		ANIMO Trauma validation sub cohort[e] (*n* = 143)		Amodels Derivation study[b] (*n* = 262)	Dmodel Derivation study (*n* = 468)
	Value[a] n (%)	*Missing* *n (%)*	*Value*[a] n (%)	*Missing* *n (%)*	*Value*[a] n (%)	*Value*[a] *n (%)*
Baseline characteristics						
Sex		7 (0.6%)		1 (0.7%)		
Female	823 (69.4%)		102 (71.8%)			182 (39%)
Male	363 (30.6%)		40 (28.2%)			
Duration current episode[c]						
Acute	420 (39.2%)	122 (10.2%)	49 (35.5%)	5 (3.5%)	262 (100%)	58 (13%)
Subacute	138 (12.9%)		11 (08.0%)			225 (48%)
Chronic	513 (47.9%)		78 (56.5%)			160 (34%)
Marital status, yes	889 (77.2%)	41 (3.4%)	102 (72.9%)	3 (2.1%)		
Currently smoking, yes	300 (25.2%)	3 (0.3%)	30 (21.0%)	0 (0.0%)		
Current medication use, yes	560 (47.1%)	3 (0.3%)	74 (51.7%)	0 (0.0%)		
Current sports, yes	783 (65.9%)	4 (0.3%)	93 (65%)	0 (0.0%)		
Disability (NDI), mean ± SD	13.0 ± 6.5	98 (8.2%)	15.9 ± 7.9	13 (9.1%)	16.5 ± 8.7	14.5 ± 6.7
Fear avoidance, FABQ scale 0–96	1053					
FABQ work subscale 0–66	26.6 ± 16.6	140 (11.7%)	30.6 ± 18.6	15 (10.5%)		
FABQ physical activity subscale	1129	64 (5.4%)	16.0 ± 14.0	8 (5.6%)		
0–30	13.4 ± 12.2		14.6 ± 7.4	10 (7.0%)		
	1103	90 (7.5%)				
	13.2 ± 7.3					
Expected recovery by patient,	1190	3 (0.3%)	143	0		
scale 1–5			57 (39.3%)			
Much better	517 (43.4%)		83 (58.0%)			
Better	662 (55.6%)		3 (02.1%)			
No change	10 (00.8%)		0 (0.00%)			
Worse	1 (00.1%)		0 (0.00%)			
Much worse	0 (0.00%)					
Dmodel for persistent neck complaints[d]						
Age, yrs.	1170					
	44.7 ± 13.7	23 (1.9%)	41.9 ± 13.8	1 (0.7%)	37.1 ± 14.2	45.4 ± 11.8
Pain, 11-point Likert scale[g]	1189					
	3.3 ± 2.7	4 (0.3%)			4.2 ± 2.1	5.7 ± 2.1
Headache, yes	707 (59.2%)		101 (70.6%)			317 (68%)
Radiating arm pain, yes	536 (44.9%)		66 (46.2%)			296 (63%)
Previous neck pain episode, yes	755 (66.9%)	64 (5.4%)	80 (59.3%)	8 (5.6%)		301(64%)
Cause of complaints trauma, yes	143 (13.0%)*	97 (8.1%)				63 (14%)
Low back pain	538 (45,1%)		65 (45.5%)			96 (21%)
Employed, yes	897 (77.1%)	29 (2.4%)	112 (79.4%)	2 (1.4%)		334 (71%)
Euro QoL 100[h]						69.9 ± 17.3
Amodel for full recovery						
NDI ≤ 32	180 (16.4%)		74 (56.9%)			
Age ≤ 35 yrs.	306 (26.2%)		49 (34.5%)			
Amodel for moderate/severe disability						
NDI ≥ 40	796 (72.7%)		40 (30.8%)			
Age ≥ 35 yrs.	888 (75.9%)		98 (69.0%)			
PDS hyperarousal subscale (0–15)[f]	481 (40.6%)	8 (0.7%)	69 (48.3%)		4.8 ± 3.8	
Outcome characteristics[i]						
Post-treatment						
Global Perceived Effect, 7-point	568	625 (52.4%)	65	78 (54.5%)		
Likert scale 0–70	129 (22.7%)		13 (20.0%)			
Completely recovered	317 (55.8%)		38 (58.5%)			
Much improved	97 (17.1%)		11 (16.9%)			
Slightly improved	25 (4.4%)		3 (4.6%)			
No change	0 (0.0%)		0 (0.0%)			
Slightly worse	0 (0.0%)		0 (0.0%)			
Much worse	0 (0.0%)		0 (0.0%)			
Worse than ever						
Disability, NDI scale 0–50	541	652 (54.7%)	64	79 (55.2%)		
	12.1 ± 11.0		8.0 ± 6.3			
Long term outcome						
Global Perceived Effect, 7-point	685	508 (42.6%)	86	57 (39.9%)		
Likert scale 0–70	157 (22.9%)		19 (22.1%)			
Completely recovered	264 (38.5%)		34 (39.5%)			
Much improved	153 (22.3%)		18 (20.9%)			

Models' performance

The ANIMO smallest outcome groups contained 122, 247, and 40 events at post-treatment for GPE, NDI recovery, and NDI moderate/severe, respectively. At long-term, these numbers were 264, 289, and 45, respectively. These numbers revealed a sufficient sample size for the Dmodel and Amodel recovery post-treatment and at long-term. The ANIMO trauma subset did not have a sufficient sample size as it contained 24 recovered people as measured by the NDI and 9 with moderate/severe outcome post-treatment, and 41 and 13 at long-term.

Discriminative performance

Models' performance measures are described in Table 3. Discriminative performance (analyzed in the trauma sub-set) of the Amodel that predicts full recovery immediately post-treatment was 0.53 (95% CI: 0.24, 0.80) and was 0.49 (95% CI: 0.26, 0.72) for long-term outcome. Discriminative performance of the Amodel that predicts ongoing moderate to severe disability post-treatment was 0.54 (95% CI: 0.40, 0.69) post-treatment and 0.54 (95% CI: 0.38, 0.69) for long- term outcome. Discriminative performance of the Dmodel was 0.53 (95% CI: 0.48, 0.58) post-treatment and 0.54 (95% CI: 0.49, 0.58) at long-term outcome. These results indicate poor discriminative performance of both models. Analysis of the Amodels in the whole ANIMO cohort at long-term follow-up revealed a discriminative performance for the model that predicts full recovery of 0.43 (95% CI: 0.40, 0.49) and for the model that predicts ongoing moderate to severe disability of 0.43 (95% CI: 0.34, 0.52), also displaying poor discriminative performance.

Table 2 (*Continued*)

	ANIMO Validation cohort (*n* = 1193)		ANIMO Trauma validation sub cohort[e] (*n* = 143)		Amodels Derivation study[b] (*n* = 262)	Dmodel Derivation study (*n* = 468)
	Value[a] n (%)	*Missing n (%)*	*Value*[a] n (%)	*Missing n (%)*	*Value*[a] n (%)	*Value*[a] *n* (%)
Slightly improved	88 (12.8%)		12 (14.0%)			
No change	12 (1.8%)		1 (1.2%)			
Slightly worse	8 (1.2%)		2 (2.3%)			
Much worse	3 (0.4%)		0 (0.0%)			
Worse than ever						
Disability, NDI scale 0—50	541 6.0 ± 5.4	515 (43.2%)	87 8.3 ± 8.0	56 (39.2%)		
Dmodel for persistent neck complaints (GPE)						
Post-treatment						
persistent complaints	122 (21.5%)		14 (21.5%)			
complete/much improved	446 (78.5%)					
Long-term						
persistent complaints	264 (38.5%)		33 (38.4%)			(43%)
complete/much improved	421 (61.5%)		51 (61,6%)			
Amodel for full recovery						
Post-treatment						
persistent complaints NDI	294 (54.3%)		51 (78.5%)			
Long term						
persistent complaints NDI	389 (57.4%)		41 (47.1%)		120 (46%)	
Amodel for moderate/severe disability						
Post-treatment						
persistent complaints NDI	40 (7.4%)		9 (14.1%)			
Long term						
persistent complaints NDI	45 (6.6%)		13 (14.9%)		69 (26%)	

Values are numbers (percentages) unless stated otherwise.
NDI = Neck Disability Index; FABQ = Fear Avoidance Beliefs Questionnaire; NRS = Numeric Rating Scale, euro QOL = Quality of Life; GPE = Global Perceived Effect; SD = Standard Deviation.
[a] Data presented as responders n (%) or mean ± SD.
[b] Complete cases of acute whiplash (*n* = 336 eligible).
[c] acute < 1 months, subacute 1—3 months, chronic >3 months.
[d] Constant and predictor's weight as Beta value.
[e] As any self-reported trauma, according to patient and/or therapist.
[f] in ANIMO Neck Bournemouth Questionnaire (NBQ) subscale ≥ 4 (how anxious, tense, uptight, irritable, difficulty concentrating/relaxing, as proxy for hyperarousal subscale of the posttraumatic stress diagnostic scale (PDS).
[g] In Dmodel studies as NRS 11-point Likert scale 0—10; in Amodel studies as VAS-scale; in ANIMO as NRS 1-point Likert scale 1—10.
[h] not available in ANIMO.
[i] Dmodel: GPE dichotomized as not complete + much improved; Amodel-moderate/severe complaints: dichotomized as NDI ≥ 30%; Amodel-full recovery: dichotomized as NDI ≤ 10%.

Calibration performance

Performance of calibration-in-the-large for the Amodel that predicts full recovery post-treatment was 0.46 (IQR: 0.13, 0.75) and 0.34 (IQR: -0.04, 0.82) for long-term outcome. The calibration slope was -0.35 (IQR: -0.57, -0.30) and -0.26 (IQR: -0.30, -0.10), respectively. For the Amodel that predicts ongoing moderate/severe disability post-treatment, calibration-in-the-large was -0.63 (IQR: -1.06, -0.08) and -1.13 (IQR: -1.76, -0.79) for long-term outcome. The calibration slope was -0,06 (IQR: 0-.12, 0.00) and -0.01 (IQR: -0.04, 0.06), respectively. The Hosmer-Lemeshow goodness of fit test was significant for both Amodels. Performance of calibration-in-the-large for the Dmodel was -0.97 (IQR: -1.03, -0.79) post-treatment and -0.33 (IQR: -0.39, -0.31) for long-term outcome. The calibration slope was -0.06 (IQR: -0.15, -0.06) and 0.23 (IQR: 0.14, 0.28), respectively. The Hosmer-Lemeshow goodness of fit test was significant for all D model outcomes.

Dmodel calibration plots are shown in Fig. 1. These values deviate substantially from the intercept of 0 and the ideal calibration slope of 1 and show poor calibration of both models.

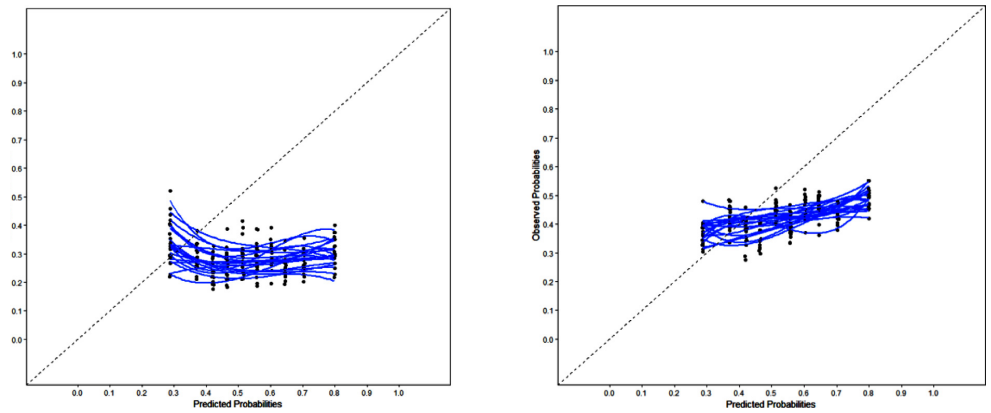


Fig. 1 Calibration plots with 20 calibration lines (blue) of each imputed dataset. Predicted probabilities are plotted against actually observed outcomes in relation to the ideal 45° line of perfect prediction (dotted line) in ANIMO decile subgroups of predicted events. Ideally, all blue lines lay exactly on the dotted line. Dmodel long-term outcome left figure, post-treatment right figure.

Sensitivity analysis

Sensitivity analyses of discriminative performance in com- plete cases demonstrated lower c-statistics of 0.36 (95% CI: 0.31, 0.41) and 0.44 (95% CI: 0.39, 0.49) for the Amodel that predicts full recovery at post-treatment and long-term, respectively. For the Amodel that predicts ongoing moderate/severe disability, these values were 0.46 (95% CI: 0.36, 0.57) and 0.42 (95% CI: 0.34, 0.52), respectively. Dmodel's discriminative performance was 0.56 (95% CI: 0.50, 0.63) and 0.54 (95% CI: 0.50, 0.69), respectively. Also, complete case analyses displayed poor discriminative performance for all models.

Discussion

External validation in a cohort of people with neck pain of a two-way WAD model (Amodel) that predicts disability measured by the NDI, and a non-specific neck pain model (Dmodel) that predicts recovery measured by the GPE, was not successful as their discriminative performance and calibration clearly did not meet expected thresholds. A third prognostic model could not be evaluated in this study because of variable discrepancy across data sets. The Amodels' discriminative performance was substantially below 0.70 for all time points. However, its discriminative and calibration performance could not be compared with the original studies because these measures were not described and our study is the first in presenting Amodels' performance measures.[18,27] The Amodel full recovery broad confidence intervals obtained in the trauma subset included AUC 0.70 values close to the upper bounds.

These broad intervals could be explained by too few events, because the ANIMO trauma subset did not reach the minimum of 100 events in the smallest outcome group. Analysis in the whole ANIMO cohort, containing sufficient events, revealed small intervals but with 0.52 as the upper bound value.

The Dmodel's discriminative performance in the original study was 0.66 (95% CI: 0.61, 0.71) at internal validation and 0.65 (95% CI: 0.59, 0.71) at external validation. Our validation study revealed a lower 0.53 (95% CI: 0.48, 0.58) AUC post-treatment and 0.54 (95% CI: 0.49, 0.58) AUC for long-term predictions. A decrease in discriminative performance from derivation to validation is not unusual.33 Dmodel's performance at development was already below our cut-off of 0.70 for AUC and a 0.12 decrease of an overfitted model in another population with a different case-mix is not an unexpected finding.

Additionally, there may be little distinction in AUC between our validation study and the development study, as the 95% CI are close together. In addition, calibration was poor for both Dmodel and Amodels. At external validation, predictions are often too extreme due to overfitting at the development phase.[44] This results in low predictions being too low and high predictions being too high, as characterized by a calibration slope smaller than 1 and indicates that the original regression coefficients were too large.[13,45,46] In addition, we believe case-mix differences could not have been responsible for models' poor performance as these differences were relatively small.
Comparison of model performance to other studies in the field is hampered: prognostic prediction models in the musculoskeletal field typically do not reach their validation phase and methodological shortcomings are common. In fact, the few models that were evaluated for external validity usually did not present model performance by means of calibration and discrimination measures.[14,17,47]

Strengths and limitations

The strength of our study is analysis in a large cohort by state-of- the-art calibration and discrimination measures. However, there are some limitations we would like to report.
First, in ANIMO, multiple independent therapists at multiple sites were used and the broad CIs derived in the large ANIMO cohort could reflect this measurement variability. Second, the validation data set had substantial missing values, which is not unusual.[48] We applied multiple imputation procedures and sensitivity analysis on complete cases that showed comparable values of the performance measures. Third, the EuroQol predictor for the Dmodel and the hyperarousal subscale predictor for the first Amodel were not available in ANIMO and may have influenced model performance. However, this impact is probably negligible considering the 0.005 bèta value for EuroQol. We believe that the NBQ anxious subscale predictor served sufficiently as a proxy for the hyperarousal subscale, thereby, the other Amodel that did not contain this predictor performed very similar. Fourth, the predicted outcomes for the Dmodel at derivation and validation were measured at 6 months and 12 months, respectively. We believe that the impact of these different outcome times is limited as the overall prognosis for neck pain and disability for 6 and 12 months appear to be similar.[49]

Implications for practice and research

Based on our findings, the clinical use of these promising models can, at present, not be advocated. We feel this is a very important message for musculoskeletal clinicians considering the numerous models that predict outcomes in neck pain that are available for clinicians without this crucial step of subsequent external validation, which could potentially lead to undesired outcomes for patients when models are implemented too early in practice. We advise clinicians to underpin their clinical reasoning process at this moment with separate prognostic factors that can be used with more confidence, such as baseline pain intensity, baseline neck disability, age, and history of musculoskeletal disorders.[50]
The low performance of the existing prognostic models indicates that important predictors may not have been included in the models' derivation process and further search for valuable model predictors is needed.

Conclusion

External validation of two promising prognostic models on neck pain recovery in primary care was not successful and their clinical use can, at present, not be advocated. Currently, no useful models are available for clinicians to predict outcomes in people with neck pain. New insights on potentially valuable prognostic factors are needed to strengthen models' derivation and updating procedures.

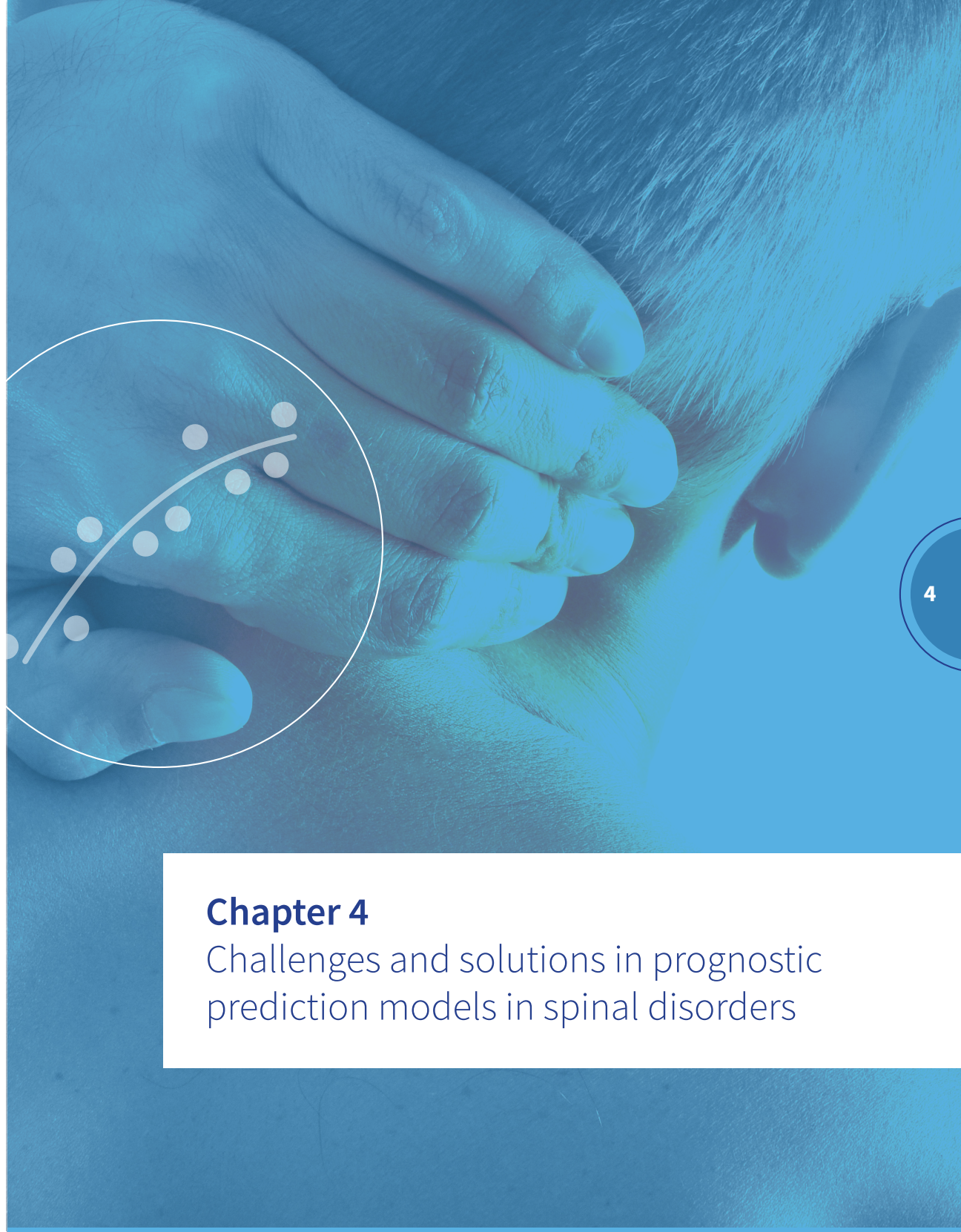
References

1. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990 2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388 (10053):1545 1602. https://doi.org/10.1016/S0140-6736(16) 31678-6.
2. Hurwitz EL, Randhawa K, Yu H, Co^te' 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. https://doi. org/10.1007/ s00586-017-5432-9. 0123456789.
3. Hush JM, Lin CC, Michaleff Z a, Verhagen A, Refshauge KM. Prog- nosis of acute idiopathic neck pain is poor: a systematic review and meta-analysis. Arch Phys Med Rehabil. 2011;92(5): 824 829. https://doi. org/10.1016/j.apmr.2010.12.025.
4. van der Velde G, Yu H, Paulden M, et al. Which interventions are cost-effective for the management of whiplash-associated and neck pain-associated disorders? A systematic review of the health economic literature by the Ontario Protocol for Traffic Injury Management (OPTIMa) Collaboration. Spine J. 2016;16(12): 1582 1597. https://doi.org/10.1016/j.spinee.2015.08.025.
5. Vincent K, Maigne J-YY, Fischhoff C, Lanlo O, Dagenais S. Systematic review of manual therapies for nonspecific neck pain. Joint Bone Spine. 2013;80(5):508 515. https://doi.org/ 10.1016/j. jbspin.2012.10.006.
6. Gross A, Kay T, Paquin J, et al. Exercises for mechanical neck disorders (Review). Cochrane Database Syst Rev. 2015(1). https://doi.org/10.1002/14651858.CD004250.pub5.Copyright.

7. Hurwitz EL, Carragee EJ, van der Velde G, et al. Treatment of neck pain: noninvasive interventions. Results of the bone and joint decade 2000-2010 task force on neck pain and its associated disorders. J Manipulative Physiol Ther. 2009;32(2 SUPPL): S141 S175. https://doi.org/10.1016/j.jmpt.2008.11.017.
8. Riley RD, Van Der Windt DA, Croft P, Moons KGM. Prognosis Research in Healthcare. First. Oxford University Press; 2019.
9. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346(feb05 1):e5595. https://doi.org/10.1136/bmj. e5595.
10. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer Science and Business Media; 2019.
11. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338(7708):1432 1435. https://doi.org/10.1136/bmj.b605.
12. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245 247. https://doi.org/10.1016/j.jclinepi. 2015.04.005.
13. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925 1931. https://doi.org/ 10.1093/eurheartj/ehu207.
14. van Oort L, van den Berg T, Koes BW, et al. Preliminary state of development of prediction models for primary care physical therapy: a systematic review. J Clin Epidemiol. 2012;65(12): 1257 1266. https://doi.org/10.1016/j.jclinepi.2012.05.007.
15. Stanton TR. Clinical prediction rules that don't hold up—where to go from here? J Orthop Sport Phys Ther. 2016;46 (7):502-505. https://doi.org/10.2519/jospt.2016.0606.
16. Beneciuk JM, Bishop MD, George SZ. Clinical prediction rules for physical therapy interventions: a systematic review. Phys Ther. 2009;89(2):114 124. https://doi.org/10.2522/ptj.20060295.
17. 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 musculoskeltal primary care: a systematic review. J Physiother. 2018;64 (1):16 23. https://doi.org/10.1016/j.jphys.2017.11.013.
18. 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(10):2198 2206. https://doi.org/10.1016/j. pain.2013.07.001.
19. Sterling M, Jull G, Vicenzino B, Kenardy J, Darnell R. Physical and psychological factors predict outcome following whiplash injury. Pain. 2005;114(1):141 148. https://doi.org/10.1016/j.pain.2004.12.005.
20. 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(17):E827 E835. https://doi. org/10.1097/BRS.0b013e3181d85ad5.
21. 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(1):55. https://doi.org/10.7326/M14-0697.
22. 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. November 2019. https://doi.org/10.1016/ j.jmpt.2019.07.001.
23. Vernon H, Mior S. The neck disability index: a study of reliability and validity. J Manip Physiol Ther. 1991;14(7):409 415.

24. Ailliet L, Rubinstein SM, de Vet HCW, van Tulder MW, Terwee CB. Reliability, responsiveness and interpretability of the neck disability index-Dutch version in primary care. Eur Spine J. 2014;24(1):88 93. https://doi.org/10.1007/s00586- 014-3359-y.
25. Schmitt M a, de Wijer A, van Genderen FR, van der Graaf Y, Helders PJ, van Meeteren NL. The neck bournemouth ques- tionnaire cross-cultural adaptation into dutch and evalua- tion of its psychometric properties in a population with subacute and chronic whiplash associated disorders. Spine (Phila Pa 1976). 2009;34(23):2551 2561. https://doi.org/ 10.1097/BRS.0b013e3181b318c4.
26. Landers MR, Creger R V, Baker C V, Stutelberg KS, Landers M, Creger R, Baker C SK. The use of fear-avoidance beliefs and nonorganic signs in predicting prolonged disability in patients with neck pain. Man Ther. 2008;13(3):239 248. https://doi. org/10.1016/j.math.2007.01.010.
27. Ritchie C, Hendrikz J, Jull G, Elliott J, Sterling M. External vali- dation of a clinical prediction rule to predict full recovery and ongoing moderate/severe disability following acute whiplash injury. J Orthop Sports Phys Ther. 2015;45(4):242 250. https://doi.org/10.2519/jospt.2015.5642.
28. Sterling M, Hendrikz J, Kenardy J, et al. Assessment and valida- tion of prognostic models for poor functional recovery 12 months after whiplash injury: a multicentre inception cohort study. Pain. 2012;153(8):1727 1734. https://doi.org/10.1016/ j.pain.2012.05.004.
29. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147 177. https://doi.org/ 10.1037/1082-989X.7.2.147.
30. 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 (2):205 214. https://doi.org/10.1016/j.jclinepi.2009.03.017.
31. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM, Der HG Van. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087 1091. https://doi. org/10.1016/j.jclinepi.2006.01.014.
32. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009;338:1 10. https://doi.org/10.1136/ bmj.b2393.
33. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691 698. https://doi.org/10.1136/heartjnl-2011-301247.
34. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172 (8):971 980. https://doi.org/10.1093/aje/kwq223.
35. Harrell FE. Evaluating the Yield of Medical Tests. JAMA J Am Med Assoc. 1982;247(18):2543. https://doi.org/10.1001/jama. 1982.03320430047030.
36. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Wiley; 2013.
37. Wynants L, Collins G, Van Calster B. Key steps and common pit- falls in developing and validating risk models. BJOG An Int J Obstet Gynaecol. 2016:1 10. https://doi.org/10.1111/1471- 0528.14170.
38. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279 289. https://doi.org/10. 1016/j.jclinepi.2014.06.018.
39. Vergouwe Y, Steyerberg EW, Eijkemans MJCC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58(5):475 483. https://doi.org/10.1016/ j.jclinepi.2004.06.017.
40. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35(2):214 226. https://doi.

org/10.1002/sim.6787.
41. Janssen KJM, Vergouwe Y, RT Da, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem. 2009;55(5):994 1001. https://doi.org/10.1373/clin- chem.2008.115345.
42. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40 49. https://doi.org/10.1002/ mpr.329.
43. 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):1 8. https://doi.org/10.1186/1471-2288-9-57.
44. RiD R, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ20162016: i3140. https://doi.org/10.1136/bmj.i3140. Under review.
45. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe, Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61 (1):76 86. https://doi.org/10.1016/j.jclinepi.2007.04.018.
46. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the perfor- mance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010;21(1):128 138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.Assessing.
47. 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. https://doi.org/10.1016/j.math.2011.05.001.
48. Ambler G, Omar RZZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. Stat Methods Med Res. 2007;16 (3):277 298. https://doi.org/10.1177/0962280206074466.
49. Henschke N, Ostelo RW, Terwee CB, van der Windt DA. Identifying generic predictors of outcome in patients presenting to primary care with non-spinal musculoskeletal pain. Arthritis Care Res (Hoboken). 2012;92(5). https://doi.org/10.1002/acr.21665.
50. Walton DM, Carroll LJ, Kasch H, 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(1):494 505. https://doi.org/10.2174/ 1874325001307010494.



Chapter 4
Challenges and solutions in prognostic prediction models in spinal disorders

Chapter 4. Challenges and solutions in prognostic prediction models in spinal disorders

Roel W. Wingbermühle, Alessandro Chiarotto, Bart Koes, Martijn W. Heymans, Emiel van Trijffel

Journal of Clinical Epidemiology. 2021 Apr; 132: 125-130

Abstract

Methodological shortcomings in prognostic modelling 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 modelling for patients with spinal disorders are presented with their potential solutions, as related to the choice of study participants, the purpose of studies, limitations in measurements of outcomes and predictors, the 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.

1. Introduction

Prediction models estimate the probability of a condition being present or a future health outcome occuring 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 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 labelled 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-term clinical course is unfavourable in a substantial proportion of people [8,9]. Prognostic models intend to distinguish patients with an unfavourable long-term course from those with a favourable course, have the potential to decrease patients' burden, and can make contributions to cost-effective healthcare. Early comprehensive treatment given to people with a favourable short-term course is unnecessary and probably not cost-effective. Interventions in this group with a favorable prognosis can even be contra-effective. 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 modelling 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 modelling 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 over-optimistic 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 healthcare [12,14].

2.2 Specific methodological challenges

2.2.1. Challenge 1: problems with the 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 a demarcation of subgroups with expected different prognosis (e.g., WAD Grade III) is recommended.

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 modelling for patients with spinal disorders as related to the choice of study participants, the purpose of studies, heterogeneity of outcome measurements, limitations in measurements of outcomes and predictors, the 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.

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 adequately 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 been 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 modelling to compare recovered (or improved) versus non recovered (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 in 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 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 modelling 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, the 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 a 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 a 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 an 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

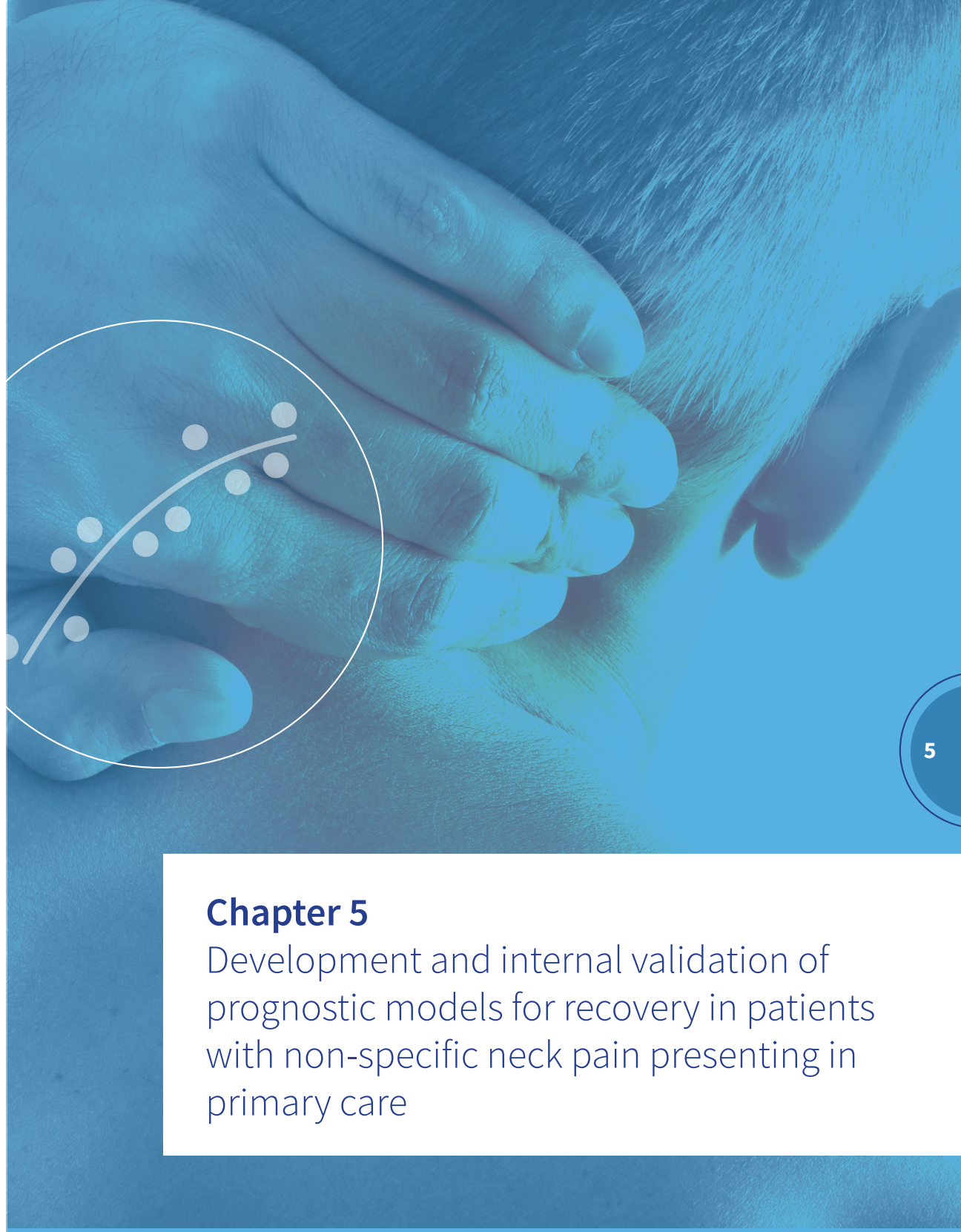
A clear description of participants’ recruitment and selection is paramount in spinal research prognostic models, with a sharp demarcation of subgroups with expected different prognosis and a clear message about the models intended use. There is a need for studies to investigate the influence and impact of measurement limitations that afflict widely used PROMs on key properties such as content validity, structural validity, and measurement error, allowing researchers to account for and correct these measurement limitations. Several problems in spinal prognostic modelling 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 modelling prognostic factors and their interactions, dynamically and in real-time, in large

linked data sets from local electronic healthcare 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 modelling 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 pre- diction 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 non-specific 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 func- tional 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 ap- proaches: three systematic reviews. NPJ Digit Med 2020; 3(1):93.
34. Schwartz JT, Gao M, Geng EA, Mody KS, Mikhail CM, Cho SK. Ap- plications of machine learning using electronic medical records in spine surgery. Neurospine 2019;16(4):643–53.



5

Chapter 5

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

Chapter 5. Development and internal validation of prognostic models for recovery in patients with non-specific neck pain presenting in primary care

Roel W. Wingbermühle, Alessandro Chiarotto, Emiel van Trijffel, Bart Koes, Arianne P. Verhagen, Martijn W. Heymans

Physiotherapy. 2021 Dec; 113: 61-72

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 a 1-year follow-up. **Results:** All post-treatment models exhibited acceptable discriminative performance after the derivation (AUC ≥ 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.

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]. Noninvasiveprimary 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 patient's prognosis; at the same time, a wait-and-see approaching 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: age, sex, baseline pain intensity, baseline disability, and 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 an adequate sample size, and it should be internally validated. This study aimed 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. The mean treatment duration was 37.9 days, the 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 the 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' expectations 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 were estimated in each of the 50 completed datasets and then combined [39,42,43].

Statistical analysis

Regression model assumptions such as the 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 datasets 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 Rversion 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 of 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 (Appendix 3) and cases with no outcome data (Table 1). The 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. The mean baseline pain intensity was 4.8 (SD 2.1) and the 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.

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.

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.

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.

Missing values

Several baseline characteristics had more than 5% missing values and a few had up to 13% (Table 1). The 1-year outcome values reached about 45% missing values and the post-treatment reached about 55% (Table 2). Baseline characteristics were comparable between complete cases (Appendix 2) and cases without outcome data. 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 since 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. Furthermore, 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 that there is no multicollinearity between variables.

Derived models

The derived models for post-treatment prediction are described in Table 3 and Appendix 1 and those for 1-year prediction are 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 (betas) 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.

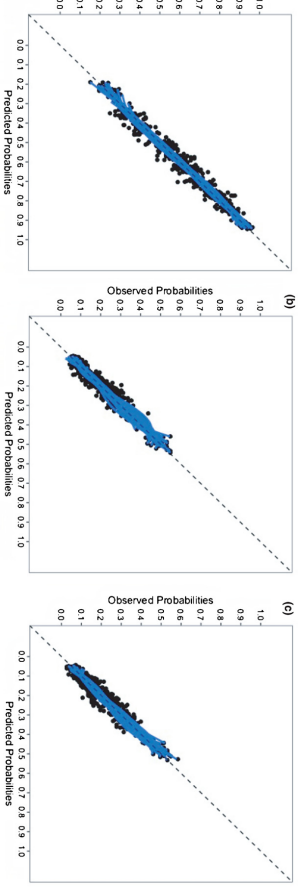


Fig 1.

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

Predictors in the models

Neck pain duration was a predictor in all models (Appendix 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

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)				
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)	0.13 [0.12 to 0.14] [$]	0.09 [0.08 to 0.11] [$]	0.70 (0.56 to 0.81) [$$]	0.67 [0.66 to 0.69] [$]
Baseline pain (NRS 0 to 10)[d]	0.19 (0.07, 0.31)	1.21 (1.07 to 1.36)				
BNQ anxiety & depression (0 to 20)[e]	0.04 (−0.00, −0.10)	1.05 (1.00 to 1.10)				
Age	0.01 (−0.00, 0.02)	1.01 (1.00 to 1.02)				
Disability model[b] **						
Constant	−2.75 (−3.58, −1.93)	0.06 (0.03 to 0.15)				
Subacute pain	0.30 (−0.27, 0.86)	1.34 (0.77 to 2.36)				
Chronic pain	0.96 (0.53, 1.40)	2.62 (1.70 to 4.03)	0.24 [0.22 to 0.26] [$]	0.21 [0.19 to 0.23] [$]	0.75 (0.63 to 0.84) [$$]	0.74 [0.72 to 0.75] [$]
Baseline disability (NDI 0 to 50[f]	0.12 (0.08, 0.16)	1.13 (1.08 to 1.17)				
Age	0.02 (0.01, 0.03)	1.02 (1.01 to 1.03)				
General sleeping problems	0.31 (−0.10, 0.72)	1.36 (0.91 to 2.05				
FABQ physical activity (0 to 24)[g]	0.02 (−0.01, 0.05)	1.02 (0.99 to 1.05)				
Perceived improvement model[c] ***						
Constant	−2.72 (−3.80, −1.64)	0.07 (0.02 to 0.19)				
Subacute pain	0.16 (−0.70, 1.03)	1.18 (0.49 to 2.81)				
Chronic pain	0.95 (0.46, 1.43)	2.57 (1.60 to 4.17)				
Low back pain	0.41 (−0.02, 0.84)	1.51 (0.98 to 2.30)	0.13 [0.11 to 0.15] [$]	0.09 [0.07 to 0.11] [$]	0.70 (0.56 to 0.80) [$$]	0.67 [0.65 to 0.69] [$]
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*).

values. The post-treatment pain model and the 1-year models derived in complete case data yielded the same or almost the same predictors (Appendix 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.

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. It 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 the 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 the 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 interviews [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 the 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 the reason of comparison with previously 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-off 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 performed 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 the 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 a 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 a 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].

Appendix 1. Prognostic models for predicting post-treatment recovery of neck pain.

* Penalized pain model (slope 0.86) is -5.94 + 0.18*Subacute pain + 0.83*Chronic pain + 0.16*Baseline pain + 0.34*BNQ-AD + 0.01*Age.

** Penalized disability model (slope 0.92) is -2.64 + 0.28*Subacute pain + 0.88* Chronic pain + 0.11*Baseline disability + 0.02*Age + 0.28* General sleeping problems + 0.02*FABQ-PA

*** Penalized perceived improvement model (slope 0.86) is -4.54 + 0.14*Subacute pain + 0.82*Chronic pain + 0.35*Low back pain + 0.03*FABQ-PA + 0.01*Age - 0.03*Baseline disability - 0.4*Previous episode + 0.33* Partaking in sporting activities

Baseline pain measured with NRS = Numeric Rating Scale (1-10-point Likert scale); NBQ-AD = Neck Bournemouth Questionnaire, Anxiety and Depression subscale (scale 0-20), sum score of 11-point numeric subscale of items 4 and 5; Baseline disability measured with NDI = Neck Disability Index (0-50 scale); FABQ-PA = Fear Avoidance Beliefs Questionnaire, Physical Activity subscale (scale 0-24); Subacute pain duration 6-12 weeks and chronic pain duration >12 weeks.

Appendix 2. Baseline characteristics and candidate model predictors of patients with non-specific neck pain for complete cases.

Baseline characteristics	Baseline for complete cases post-treatment #	Baseline for complete cases at 1y #
Age (years), Mean (SD)	46.7 (13.6) 46.3 (13.7) 46.5 (13.7)	45.3 (13.7) 45.1 (13.4) 45.2 (14.0)
Gender Female sex, n (%)	70 70 70	70 69 69
Previous neck pain episode yes, n (%)	71 71 71	67 67 67
Neck pain duration Acute 0-6 weeks, n (%) Subacute 6-12 weeks, n (%) Chronic >12 weeks, n (%)	 38 38 38 14 14 14 49 49 48	 39 39 36 13 13 13 48 48 51
Pain intensity (*NRS, scale 1-10)* [1], Mean (SD)	4.8 (2.1) 4.8 (2.1) 4.8 (2.0)	4.8 (2.1) 4.8 (2.1) 4.7 (2.1)
Disability (*NDI, scale 0-50)* [2], Median [IQR]	12.0 [8.0-17.0] 12.0 [8.0-17.0] 12.0 [8.0-17.0]	12.0 [8.0-17.0] 12.0 [8.0-17.0] 12.0 [8.0-17.0]
Accompanying headache yes, n (%)	61 61 61	58 59 59
Accompanying low back pain yes, n (%)	48 48 48	46 47 47
Accompanying radiating arm pain yes, n (%)	46 46 46	44 44 42
Accompanying general sleeping problems yes, n (%)	29 29 29	28 29 28
Smoking status yes, n (%)	24 24 24	24 24 23
Fear-avoidance believes (*FABQ-PA, scale 0-24*) [a], Median [IQR]	10.0 [5.0-14.8] 10.0 [5.0-15.0 10.0 [5.0-14.8]	11.0 [6.0-15.0] 11.0 [6.0-15.9] 11.0 [6.0-5.0]
Emotional functioning (*NBQ-AD, scale 0-20)* [b], Median [IQR]	7.0 [3.0-10.0] 7.0 [3.0-10.0]	6.0 [3.0-10.0] 6.0 [3.0-10.0]
	7.0 [3.0-10.0]	6.0 [3.0-10.0]
Partaking in sporting activities yes, n (%)	65 65 65	67 67 65
Patients' expectation to change due to treatment Much better, n (%) Better, n (%) No change, n (%) Worse, n (%) Much worse, n (%)	 42 42 42 58 58 58 1 1 1 0 0 0 0 0 0	 44 44 43 56 55 56 1 1 1 0 0 0 0 0 0

% rounded up to closest integer
[1] NRS = Numeric Rating Scale
[2] NDI = Neck Disability Index
[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
For Global Perceived Effect (GPE), NRS and NDI, respectively

Appendix 3. Predictors in the models for predicting recovery of non-specific neck pain obtained for each outcome at post-treatment and at 1-year follow-up.

Outcome	Post-treatment	At 1-year
Pain intensity	Pain duration [1] Baseline pain [1] Age [2] BNQ-AD [1, b]	Pain duration [3] Baseline pain[1] Headache [2] Sleeping problems [1] Previous episode [1] Sleeping problem [1] Low back pain [2]
Disability	Pain duration [1] Baseline disability [1] Age [1] Sleeping problems [1] FABQ-PA[1, a] Sporting [3] Previous episode [3]	Pain duration [1] Baseline disability [1] Headache [1] Age [1]
Perceived improvement	Pain duration [1] Baseline disability [2] Age [2] Low back pain [1] FABQ-PA[1, a] Partaking sporting [2] Previous episode [2]	Pain duration [1] Baseline disability [1] Headache [1] Low back pain [1] Sleeping problems [1] Low back pain [1] Female gender [1]

[1] For both imputed data model and complete case data model
[2] For imputed data models only
[3] For complete case data models only
[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).

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, Mac- dermid 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

17. P. Verhagen. External validation study of three promising models for prediction of neck pain recovery. Submitted n.d. 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 non-specific 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, Han- cock 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 logis- tic 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 systematic 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 clini- cal 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, Skill- gate 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.

6

Chapter 6

External validation and updating of prognostic models for predicting recovery of disability in people with (sub)acute neck pain was successful: broad external validation in a new prospective cohort

Chapter 6. External validation and updating of prognostic models for predicting recovery of disability in people with (sub)acute neck pain was successful: broad external validation in a new prospective cohort

Roel W. Wingbermühle, Alessandro Chiarotto, Emiel van Trijffel, Martijn S. Stenneberg, Ronald Kan, Bart W. Koes, Martijn W. Heymans.

Submitted.

Abstract

Question: Can existing prognostic models for neck pain recovery be externally validated and updated at 6- and 12-week follow-up, and post-treatment? **Design:** External validation and model updating in a new prospective cohort of three previously developed prognostic models. **Participants:** People with (sub)acute neck pain, registered for primary care physiotherapy treatment. **Outcome measures:** Recovery of disability, pain, and perceived recovery at 6 and 12 weeks, and post-treatment. **Results:** Discriminative performance of the disability model at 6 weeks was 0.73 (0.69-0.77) and reasonably well calibrated after intercept recalibration. The disability model at 12 weeks and at post-treatment showed discriminative performance values just below 0.70 and was well calibrated. Pain models and perceived recovery models did not reach acceptable performance. Cervical mobility added significant value to the disability models and pain catastrophising to the disability and pain models at 6 weeks. **Discussion:** Broad external validation of the disability model was successful in people with (sub)acute neck pain and clinicians may use this model in clinical practice with reasonable accuracy. We advise further research to assess the disability model's clinical impact, generalisability, and the quest for additional valuable model predictors.

Introduction

Neck pain is common and it remains one of the leading causes of disability in most countries.[1 2] Its burden is likely to increase even further warranting high need for rehabilitation services in primary care. [3 4] Identification at intake of people with neck pain that are unlikely to recover enables personalised care and supports the improvement of health outcomes with potential to reduce its burden. Recovery from acute neck pain mainly takes place in the first few weeks. Otherwise prognosis becomes worse potentially leading to persistent pain and disability. [5 6] Prognostic factors for predicting neck pain recovery have, more or less, been established. [7 8] However, individual factors cannot provide sufficient information to be used for accurate individualised outcome predictions.

Prognostic prediction models (further: prognostic models) provide a personalised evidence-based approach, by combining multiple predictors simultaneously to estimate a patient's future individual outcomes (e.g. neck pain intensity or neck pain related disability). [9 10] Several prognostic models for neck pain have been developed. However, methodological shortcomings are common (e.g., small sample size, no overfitting correction, lack of reporting key performance measures, predictor and outcome measurement limitations) and very few models have been externally validated. [11 12] Recently, a model for people with neck pain was developed and internally validated in a Dutch cohort of people treated with manual therapy, predicting post-treatment recovery of disability with good discriminative performance. [13] This disability model may have good potential to inform primary care clinicians about the individual prognosis of people with neck pain after treatment. However, model's broad external validation is a crucial step before they can be advocated for clinical use. [14] In fact, there should be an ongoing process of model validation and updating.[15 16] The other developed models for recovery of pain and perceived improvement did not meet commonly used thresholds for discriminative performance criteria, however, they still exhibited reasonable performance and may benefit from model updating. [13]
Therefore, the research question of this study was: Are existing post-treatment prognostic models for predicting neck pain recovery, primarily in terms of disability, secondarily in terms of pain and perceived improvement, externally valid at 6- and 12-week follow-up, and post-treatment, in a new Dutch cohort of people with neck pain treated with guideline-based usual care physiotherapy?

Method

An external validation study of three internally validated models for recovery of neck pain was performed. The research protocol of this study was registered on March 20, 2021, at https://osf.io/a6r3k/. This study is reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) recommendations. [17]
The models that were externally validated were the following: (1) a disability model (DModel): -2.64 + 0.28*Subacute pain + 0.88*Chronic pain + 0.11*Baseline disability + 0.02*Age + 0.28*Sleeping problem + 0.02*FABQ-PA; AUC 0.74 (0.72-0.75), R^2 0.21 (0.19-0.23); (2) a pain model (PModel) -5.94 + 0.18*Subacute pain + 0.83*Chronic pain + 0.16*Baseline pain + 0.34*BNQ-AD + 0.01*Age; AUC 0.67 (0.66-0.69), R^2 0.09 (0.08-0.11); and (3) a perceived improvement model (PIModel): 4.54 + 0.14*Subacute pain + 0.82*Chronic pain + 0.35*Low back pain + 0.03*FABQ-PA + 0.01*Age - 0.03*Baseline disability - 0.4*Previous episode + 0.33*Sporting activities; AUC 0.67 (0.65-0.69), R^2 0.09 (0.07-0.11).

Development cohort

The models were previously developed in ANIMO data, a cohort from 2007-2009 where manual therapists in the Netherlands recruited 1311 consecutive people between 18 and 80 years presenting with non-specific neck pain of any duration, with or without arm pain. Participants received usual care manual therapy. Study details have been described elsewhere. [18]

Validation cohort

For external validation, data from the PRONEPA cohort, which ran from November 2020 to April 2021, with a 12-week follow-up was used. PRONEPA is a prospective cohort study (registered at https://osf.io/u8rnw/ ethics committee permission METCZ20200178) that primarily aimed to evaluate prognostic factors that predict the development of chronic neck pain in people with (sub)acute neck pain (< 12 weeks), with or without radicular symptoms, registered for physiotherapy treatment. PRONEPA 2020-2021 included a convenience sample of 586 participants with neck pain, recruited by 102 physiotherapists who were graduating from a Master of Science program in manual therapy at SOMT University of Physiotherapy, Amersfoort (the Netherlands). Participants' characteristics and models' predictors were collected by the physiotherapist at baseline and at 6 weeks, and models' outcomes (Neck Disability Index [NDI], Numeric Rating Scale [NRS], Global Perceived Effect [GPE]), at 3, 6, 12 weeks, 6 months, and at post-treatment (only for NDI, NRS). Inclusion criteria were primary complaints of neck pain grade 1, 2 or 3 according to the Neck Pain Task Force [19], 18 years of age or older, minimum of three days to maximum of 12 weeks of neck pain. Exclusion criteria were past or actual cervical fractures, congenital disorders affecting cervical functioning, systemic diseases or neurological disorders affecting cervical functioning, past or actual malignant diseases, and past cervical surgery.

Validation procedure

We described and compared case mix differences (i.e., participants' characteristic values and outcome occurrence) and study characteristics (i.e., recruitment period, setting, inclusion/exclusion criteria, treatment) between the development and the validation cohorts, and we tested models' performance in the validation cohort by examining discrimination, calibration, and overall performance measures. We checked a priori, at each follow-up of the validation cohort, the number of events in the recovered and non-recovered disability, pain, and perceived improvement outcome groups for a minimum of 100 to 200 events, as advised for validation studies that predict binary outcomes. [20] We performed external validation at 6 weeks, 12 weeks, and at post-treatment (if < 12 weeks) and additionally evaluated if the models could be updated, by adding additional potential predicting variables.

Outcomes

The definition of recovery used was identical to the development study. [13] Recovery of disability, pain, and perceived recovery, were dichotomised as NDI < 8% [0-50 scale range, transformed to % by multiplying with factor 2), NRS ≤ 2 [11-point Likert scale], and GPE response options "very much better" or "much better" [7-point Likert scale]; non-recovery being their inverses), respectively.

Comparison of study characteristics

Both cohorts were recruited by students graduating from a Master of Science program in manual therapy at the same institution. Usual care physiotherapy treatment was provided in both studies. The manual therapists in the development study had more work experience (mean work experience 19.3 years, SD 7.1) and qualified manual skills compared to the physiotherapists in the validation study (mean work experience 5.4 years, SD 4.7), who were last-year Master of Science in manual therapy students. Therefore, the manual therapists may have added high-velocity thrust techniques and specific joint mobilisation techniques. From the manual therapy students, it can be expected that they have been provided care according to current Dutch guidelines. [21] Both studies excluded red flags and included people over 18 years with non-specific neck pain with or without arm pain, and with or without trauma.

The validation cohort contained n=10 (1.7%) people above 80 years, and the development study excluded people above 80 years. The development study included people with neck pain of any duration, the validation study included people with minimum neck pain of three days to a maximum of 12 weeks. For model updating, additional history, physical examination, and psychosocial variables were available in the validation cohort.

Data analysis

Missing data

We described missing predictor and outcome values and analysed the data to assume the missing data mechanism (Little's test, t-test, chi-square test, logistic regression analysis) to decide if multiple imputation (MI) was needed.

Statistical validation of models' performance

We tested linearity and model assumptions and compared observed outcomes to those predicted by the models in the validation cohort in terms of discrimination and calibration measures. [9] We calculated the model's linear predictors (lp) and individual probability of recovery for disability, pain, and perceived improvement as $p\,(y=1) = 1/(1+e^{-lp})$ for all participants at 6 weeks and 12 weeks follow-up, and at post-treatment. [22] We estimated the model's overall performance using Nagelkerke's R^2 and Brier scores.

Discriminative performance

Discriminative performance indicates whether a model can distinguish between people with neck pain with and without recovery. It was calculated as the concordance (c) statistic which is comparable to the area under the curve (AUC) of the Receiver Operating Characteristic curve (ROC) for binary data. [9] We a priori considered discriminative performance acceptable if the AUC was ≥ 0.70. [23]

Calibration performance

Calibration performance refers to the agreement between a model's predicted risks and observed outcomes. [24] We performed calibration-in-the large and present models' calibration slopes and calibration plots. [24] The models were re-estimated in the validation cohort using the linear predictor (lp) and model: logit (y)=a + b x lp. [9 24 25] We tested calibration as deviation from the ideal calibration slope of 1 and the intercept of 0 using the model with an offset procedure. Calibration plots' probabilities were calculated to allow observation if all decile groups closely fit the perfect 45° line of identity. [9 24] We performed statistical analyses using IBM SPSS 27.0 and R 2021.09.01.

Models updating

We evaluated if models' updating enhanced model performance through adjustment of the models' intercept using the calibration intercept, and models' regression coefficients using the calibration slope. [26 27 28]

Additional variables were available in the validation cohort, and we tested if a limited number of potential predictors improved the models. First, from physical examination, cervical mobility and anterior muscles endurance were used for updating the models as interventions aimed at improving these functions are effective. [29 30] Cervical mobility was measured in degrees by a total sum score of flexion, extension, and both rotations (ROMmean) using the mobile phone application Goniometro. Smartphone applications measuring spinal ROM are reliable and their clinical use is supported. [31] Measurement error of the Goniometro application using the CROM-device as reference appeared small. [32] Anterior muscle endurance was measured by the neck flexor endurance test (NFET). [33] However, it reveals only a substantial intra-and interrater reliability and a large standard error of measurement of ≥ 14.57 seconds and a minimal detectable change of 40 seconds. [34] Furthermore, pain catastrophising was considered as a potential additional predictor. [35] Catastrophising is considered a predictor for persistent pain and disability in people with chronic musculoskeletal pain and in people with whiplash-related pain. [36 37]

Pain catastrophising was measured with the pain catastrophising scale (PCS), which is a reliable and valid instrument for measuring catastrophic thinking related to pain.[38 39]

We evaluated if the models improved after including these potential candidate predictors significantly ($p < 0.157$) and enhance models' performance. [15 27 28]

Results

The predictor and outcome characteristics between the validation and derivation cohort are presented in Table 1. Due to the difference in neck pain duration inclusion criteria, the validation cohort displayed no people with chronic neck pain and 40.2% more people with acute neck pain. There were no clinically meaningful differences between the other predictors (Table 1). The amount and percentage of non-recovered people with neck pain in the validation cohort decreased from 6 to 12 weeks for all outcomes. The percentage of post-treatment non-recovered people with neck pain between the validation and derivation study is comparable for pain and differs for disability (post-treatment perceived recovery was not registered in the validation study). The number of recovered and non-recovered events in the validation cohort turned out between the required minimum of 100 to 200 events for all follow-up periods; for disability, the number of events exceeded 200 for all follow-up periods.

Table 1. Predictor and outcome characteristics of participants in the validation cohort and the development study.

	Validation cohort (n=586)	**Derivation study (n=1193)**
	Value	***Value***
Predictor characteristics		
Age (years), Mean (SD)	44.0 (15.7)	44.7 (13.7)
Gender Female sex, n (%)	393 (67.1)	823 (69)
Previous neck pain episode yes, n (%)	490 (83.6)	755 (67)
Neck pain duration Acute 0-6 weeks, n (%) Subacute 6-12 weeks, n (%) Chronic >12 weeks, n (%)	464 (79.2) 122 (20.8)	420 (39) 138 (13) 513 (48)
Pain intensity (*NRS, scale 0-10, scale 1-10 resp.)* [1], Median [IQR], Mean (SD)	6.0 [4-7]	4.8 (2.1)
Disability (*NDI, scale 0-50)* [2], Median [IQR]	11.0 [8-16]	12.0 [8-17]
Accompanying low back pain yes, n (%)	167 (28.5)	538 (45)
Accompanying general sleeping problems yes, n (%)	280 (47.8)	337 (28)
Fear-avoidance believes (*FABQ-PA, scale 0-24*) [a], Median [IQR]	8.0 [4-13]	11.0 [6-15]
Emotional functioning (*NBQ-AD, scale 0-20)* [b], Median [IQR]	5.0 [2-9]	7.0 [3-10]
Partaking in sporting activities yes, n (%)	404 (68.9)	783 (66)
Potential predictors characteristics		
Mean range of motion (degrees, sum score), Mean (SD)	62.3 (11.0)	
Neck flexor muscle endurance (seconds), Median [IQR]	30.5 [21-46]	
Pain catastrophizing scale (PCS, scale 0-52), Median [IQR]	6.0 [2-12]	
Outcome characteristics		
6 weeks		
Pain intensity 6 weeks (*NRS, scale 0-10, scale 1-10 resp.)* [1], Median [IQR] Not recovered, n %	2.0 [1-4] 220 (38.1)	
Disability 6 weeks (*NDI, scale 0-50)* [2], Median [IQR] Not recovered, n %	5 [2-9] 349 (60.4)	
Global perceived improvement (*GPE, 7-point Likert scale*) [3] Completely recovered Much improved Slightly improved No change Slightly worse Much worse Worse than ever Not recovered, n %	 127 (22) 277 (48) 132 (23) 31 (5) 8 (1) 3 (1) 0 (0) 174 (30.1)	
12 weeks		
Pain intensity 12 weeks (*NRS, scale 0-10, scale 1-10 resp.)* [1], Median [IQR] Not recovered, n %	1 [0-3] 167 (28.6)	
Disability 12 weeks (*NDI, scale 0-50)* [2], Median [IQR]	3 [25-75]	

Not recovered, n %	272 (46.7)	
Global perceived improvement (*GPE, 7-point Likert scale*) [3]		
Completely recovered	201 (35)	
Much improved	232 (40)	
Slightly improved	101 (17)	
No change	31 (6)	
Slightly worse	13 (2)	
Much worse	2 (0)	
Worse than ever	2 (0)	
Not recovered, n %	150 (25.7)	
Post-treatment		
Pain intensity (*NRS, scale 0-10, scale 1-10 resp.)* [1], Median [IQR]	1 [0-2]	2.0 [1-2]
Not recovered, n %	134 (23.1)	112 (21)
Disability (*NDI, scale 0-50)* [2], Median [IQR]	3 [1-7]	5.0 [1-9]
Not recovered, n %	274 (47.5)	290 (58)
Global perceived improvement (*GPE, 7-point Likert scale*) [3]		
Completely recovered		127 (24)
Much improved		287 (55)
Slightly improved		83 (16)
No change		24 (5)
Slightly worse		0 (0)
Much worse		0 (0)
Worse than ever		0 (0)
Not recovered, n %		107 (21)

% Rounded up to closest integer
[1] NRS = Numeric Rating Scale
[2] NDI = Neck Disability Index
[3] GPE = Global Perceived Effect
[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

Models' performance

Missing data

The number of variables with missing data and the amount of missing data was very low with the vast majority between 0 to 1%. Six variables had just a little more than 1% missing values: the cervical mobility predictor 1.2%, the three outcome variables at 6 weeks 1.4%, post-treatment pain 1.2%, and post-treatment disability 1.5%. There is little gain from MI for these low proportions of missing data. [40] In addition, after analysing the missing data and checking the researcher's logbooks for reasons of missing outcomes, we assumed that the Missing Completely at Random missingness was plausible. Consequently, we decided there was no need for multiple imputation and complete case analysis was acceptable.

Models' performance measures before updating

We tested and evaluated linearity and concluded that non-linear transformation would not be advantageous. [41] Models' validation performance is described in Table 2. The disability model at 6 weeks showed acceptable discriminative performance with a c-statistic equal to 0.73 (95% CI: 0.69-0.77).

The disability models at 12 weeks and post-treatment showed discriminative performance values of 0.69 (95% CI: 0.64-0.73) and 0.68 (95% CI: 0.63-0.72), respectively. The pain models and perceived improvement models did not reach acceptable levels of the performance measures (Table 2). Calibration curves are displayed in Figure 1.

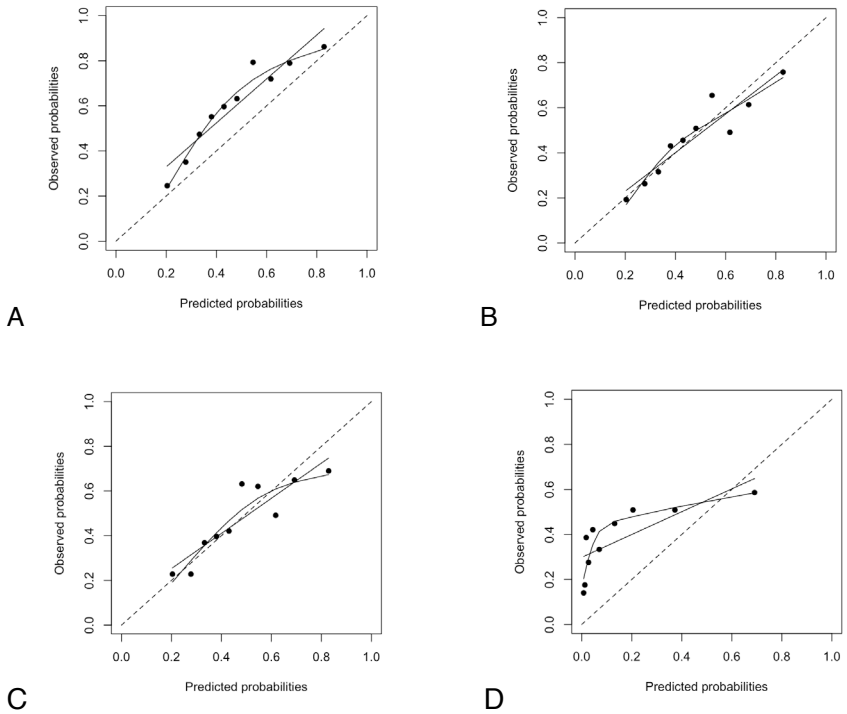


Figure 1 Models' calibration curves at validation

Disability model calibration curve at 6 weeks (A), at 12 weeks (B), at post-treatment (C).

Pain model calibration curve at 6 weeks (D), at 12 weeks (E), at post-treatment (F).

Perceived improvement model calibration curve at 6 weeks (G), at 12 weeks (H).

	Discrimination c-statistic*	**N-R²** #	**Brier Score**	**Testing Calibration In-the-large** (intercept)	**Testing Calibration** slope
Disability Model					
6 weeks	0.73 (0.69-0.77)	0.20	0.20	0.60 ##	1.10
12 weeks	0.69 (0.64-0.73)	0.14	0.22	-0.06	0.83 ## $
Post-treatment **	0.68 (0.63-0.72)	0.12	0.23	-0.05	0.76 ## $
Pain Model					
6 weeks	0.66 (0.62-0.71)	0.10	0.22	0.29	0.32 ##
12 weeks	0.66 (0.61-0.71)	0.09	0.19	-0.16	0.31 ##
Post-treatment **	0.61 (0.56-0.67)	0.04	0.17	-0.69 ##	0.21 ##
Perceived Improvement Model					
6 weeks	0.53 (0.48-0.58)	0.00	0.21	-1.91	0.21
12 weeks	0.54 (0.48-0.59)	0.00	0.19	-2.59	0.31

* As logit with 95% low and 95% up
** if <12 weeks
# Nagelkerke's R^2
significant deviation (intercept from 0, slope from 1) for test LP fit
$ not-significant deviation for test intercept and slope separate with offset procedure

Model updating

We assessed model updating for the disability models and pain models. Based on the model performance, we decided that further testing of the perceived improvement models was not useful. If intercept and/or slope values differed significantly after testing with the logit (y)=a + b x lp offset procedure, we updated the models with the values found, and subsequently re-evaluated models' performance. The calibration performance of the disability model at 6 weeks clearly improved from intercept correction using the found 0.6 value (Figure 2A), the discriminative performance did not change after this correction and remained at the same acceptable performance of c-statistic of 0.73 (95% CI: 0.69-0.77).
The other models' calibration performance did not improve, and discrimination remained identical. For further testing, we used the recalibrated 6 weeks disability model and did not adjust the other models.
Testing the additional variables revealed that the cervical mobility variable and pain catastrophising variable added significantly ($p < 0.157$) to the 6 weeks recalibrated disability model. The cervical mobility variable added significantly to the 12 weeks and post-treatment disability models. The pain catastrophising variable added significantly ($p < 0.157$) to the 6 weeks pain model. The neck flexor endurance test variable showed no additional significant value for the models.

We included the significantly adding variables and their weights to the disability models at the three follow-ups and the pain model at 6 weeks, and re-evaluated models' performance. Adding cervical mobility and pain catastrophising variables to the 6 weeks recalibrated disability model, slightly improved discrimination to 0.74 (95% CI: 0.70-0.78). Adding cervical mobility to the 12 weeks and post-treatment disability models showed c-statistics of 0.69 (95% CI: 0.65-0.73) and 0.69 (95% CI: 0.65-0.73), respectively. The calibration performance of all the disability models initially was overfitted and recovered after intercept recalibration (Figure 2B, C, D). The discrimination and calibration performance of the 6 weeks pain model did not improve.

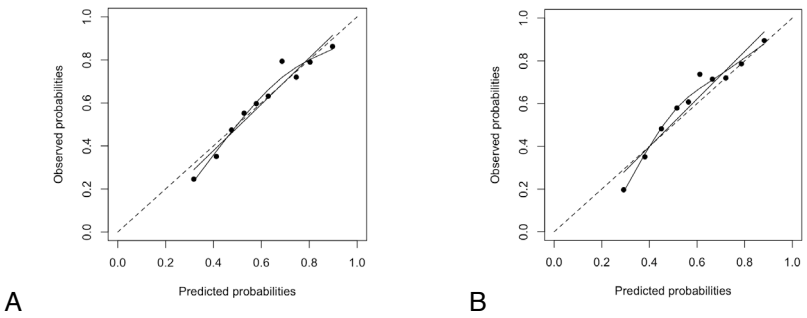


Figure 2 Models' calibration curves after updating

Adjusted disability model calibration curve at 6 weeks, after recalibration with the 0.6 calibration intercept (A), after adding the cervical mobility and pain catastrophising predictors and recalibration with the intercept (B). Disability model calibration curve at 12 weeks, after adding the cervical mobility predictor and recalibration with the intercept (C). Disability model calibration curve at post-treatment, after adding the cervical mobility predictor and recalibration with the intercept (D).

Discussion

The disability model for prediction of neck pain recovery remained discriminatory at 6 weeks in a different, external cohort of people with neck pain, coming from an independent physiotherapy setting with a different case mix. At 12 weeks and post-treatment, it showed nearly acceptable performance (c-statistics of 0.69 (95% CI: 0.64-0.73) and 0.68 (95% CI: 0.63-0.72), respectively). The pain model and perceived improvement model could not be externally validated, which was expected since internal validation was not acceptable as well. [13] Cervical mobility added significant value to the disability model at all follow-up periods and pain catastrophising also to the 6 weeks pain model. Model updating hardly affected discriminative and overall performance, whereas the different levels of updating were reflected in the shape of the calibration curves. The additional predictors improved model performance minimally and may have insufficient gain to be used clinically for purely prognostic purposes.

Few prognostic models for recovery of non-specific neck pain have been exposed to external validation. [11 42] Until now, no non-specific neck pain model has been successfully externally validated with reporting of both discrimination and calibration performance measures as recommended by TRIPOD. [17] One model stood out as it was evaluated in several external validation studies, whereby all these studies reported AUC values below 0.70. [43 44 45]
The strength of this broad external validation study is that it was conducted in a cohort with sufficient power with very few missing values. A model is more challenged in broad than small external validation, indicating a better test for its generalisability. [46]

This disability model keeps performing well at a different follow-up period in a cohort of people with neck pain with a different case mix, especially regarding the duration of pre-existing neck pain complaints. In addition, participants were treated recently, reflecting current physiotherapy guidelines.
We needed to recalibrate the 6 weeks disability model which is often needed in validation studies and indicates a difference in baseline risk between the development and validation study that was not reflected by the model predictors. This could be explained by the difference in non-recovery percentage for disability. [26 28 47] Eyeballing the disability models' calibration slopes revealed that some group mean values still were somewhat scattered around the perfect line of identity. This scattering may have been less if we had decided that non-linear transformation was advantageous, at the expense of clinical manageability. Furthermore, the use of predictor weights gained by fitting the models anew may have improved model's predictive performance. However, this implies model revision and subsequent external validation and is not preferred over simple recalibration. [28]

We advise further research to assess the disability model's clinical impact. Additional external validation studies in another clinical context (e.g., other countries, other healthcare providers and settings) may add knowledge to the model's generalisability. Furthermore, model's relatively low explained variance indicates there are still predictors for non-recovery missing and the quest for additional valuable predictors continues. Additionally, it may be of interest to further evaluate the cervical mobility and pain catastrophising predictors.

Although they showed minimal impact on prognostic performance in this study, being modifiable factors, they may have predictive capacity depending on specific treatments. For instance, the cervical mobility predictor may show predictive capacity depending on mobilisation treatment, and the prognostic effect of the pain catastrophising predictor may depend on cognitive-behavioural therapy.

Broad external validation of the disability model was successful and this model is generalisable to current physiotherapy settings and can be used in clinical practice with reasonable confidence. We advocate for physiotherapists to use the disability model at intake for the prognosis of people with neck pain to assist in clinical decisions concerning the recovery of neck pain disability at 6 weeks. We advise further research to assess the disability model's clinical impact and generalisability.

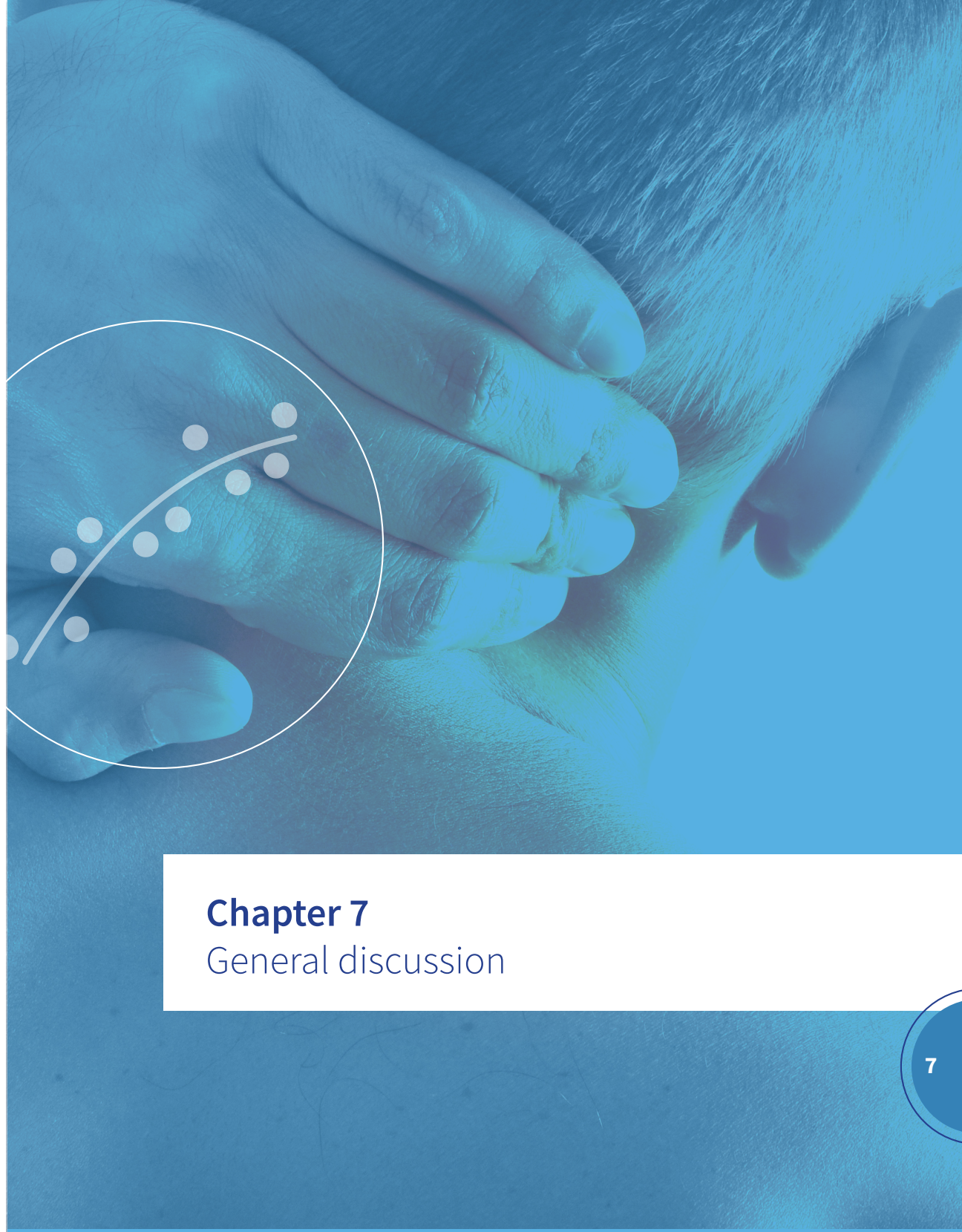
What is already known on this topic: Clinical use of currently published models for predicting recovery of non-specific neck pain cannot be advised.

What this study adds: A model for predicting recovery of disability at 6 weeks in people with neck pain was successfully broad externally validated and is advised for use in clinical practice.

References

1. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1545-1602. doi:10.1016/S0140-6736(16)31678-6
2. Abajobir AA, Abate KH, Abbafati C, 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(10100):1260-1344. doi:10.1016/S0140-6736(17)32130-X
3. Safiri S, Kolahi AA, Hoy D, et al. Global, regional, and national burden of neck pain in the general population, 1990-2017: Systematic analysis of the Global Burden of Disease Study 2017. BMJ. 2020;368. doi:10.1136/bmj.m791
4. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396(10267):2006-2017. doi:10.1016/S0140-6736(20)32340-0
5. Hush JM, Lin CC, Michaleff Z a, 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):824-829. doi:10.1016/j.apmr.2010.12.025
6. 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(8):1237-1244. doi:10.1016/j.pain.2013.03.032
7. Verwoerd M, Wittink H, Maissan F, de Raaij E, Smeets RJEM. Prognostic factors for persistent pain after a first episode of nonspecific idiopathic, non-traumatic neck pain: A systematic review. Musculoskelet Sci Pract. 2019;42(March 2018):13-37. doi:10.1016/j.msksp.2019.03.009
8. 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(1):e012901. doi:10.1136/bmjopen-2016-012901
9. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer Science and Business Media,; 2019.
10. Riley RD, Van Der Windt DA, Croft P, Moons KGM. Prognosis Research in Healthcare. first. Oxford University Press; 2019.
11. 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. doi:10.1016/j.jphys.2017.11.013
12. Wingbermühle RW, Chiarotto A, Koes B, Heymans MW, van Trijffel E. Challenges and solutions in prognostic prediction models in spinal disorders. J Clin Epidemiol. 2021;132:125-130. doi:10.1016/j.jclinepi.2020.12.017
13. Wingbermühle RW, Chiarotto A, van Trijffel E, Koes B, Verhagen AP, Heymans MW. Development and internal validation of prognostic models for recovery in patients with non-specific neck pain presenting in primary care. Physiotherapy. 2021;113:61-72. doi:10.1016/j.physio.2021.05.011
14. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-247. doi:10.1016/j.jclinepi.2015.04.005
15. Steyerberg E, Moons KGM, van der Windt D, et al. Prognosis research strategy (PROGRESS) series 3xxd: prognostic models. Br Med J. 2012;10(2). doi:10.1371/jour-
16. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279-289. doi:10.1016/j.jclinepi.2014.06.018
17. HAYASHI K. Pcm Stereo Recorder. NHK Lab Note. 1970;7594(134):1-9. doi:10.1136/bmj.g7594
18. 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(7):471-479. doi:10.1016/j.jmpt.2019.07.001
19. Guzman J, Hurwitz EL, Carroll LJ, 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. J Manipulative Physiol Ther. 2009;32(2 SUPPL.):17-28. doi:10.1016/j.jmpt.2008.11.007
20. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. Stat Med. 2016;35(2):214-226. doi:10.1002/sim.6787
21. Bier JD, Scholten-Peeters WG., Staal JB, et al. Clinical Practice Guideline for Physical Therapy Assessment and Treatment in Patients With Nonspecific Neck Pain. Phys Ther. 2018;98(3):162-171. doi:10.1093/ptj/pzx118
22. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172(8):971-980. doi:10.1093/aje/kwq223
23. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Wiley; 2013.
24. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207

25. 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(2):205-214. doi:10.1016/j.jclinepi.2009.03.017
26. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. Stat Med. 2017;36(28):4529-4539. doi:10.1002/sim.7179
27. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-698. doi:10.1136/heartjnl-2011-301247
28. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61(1):76-86. doi:10.1016/j.jclinepi.2007.04.018
29. Coulter ID, Crawford C, Vernon H, 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(2):E55-E70. http://www.ncbi.nlm.nih.gov/pubmed/30921975.
30. Martin-Gomez C, Sestelo-Diaz R, Carrillo-Sanjuan V, Navarro-Santana MJ, Bardon-Romero J, Plaza-Manzano G. Motor control using cranio-cervical flexion exercises versus other treatments for non-specific chronic neck pain: A systematic review and meta-analysis. Musculoskelet Sci Pract. 2019;42(September 2018):52-59. doi:10.1016/j.msksp.2019.04.010
31. Nuhmani S, Khan MH, Kachanathu SJ, Bari MA, Abualait TS, Muaidi QI. Reliability and validity of smartphone applications to measure the spinal range of motion: A systematic review. Expert Rev Med Devices. 2021;18(9):893-901. doi:10.1080/17434440.2021.1962290
32. Verstelle K. Reliability of the cervical flexion rotation test and the C0-C2 axial rotation test in an elderly patient population. Master thesis. 2021.
33. Harris KD, Heer DM, Roy TC, Santos DM, Whitman JM, Wainner RS. Reliability of a Measurement of Neck Flexor Muscle Endurance. Phys Ther. 2005;85(12):1349-1355. doi:10.1093/ptj/85.12.1349
34. Juul T, Langberg H, Enoch F, Søgaard K. The intra- and inter-rater reliability of five clinical muscle performance tests in patients with and without neck pain. BMC Musculoskelet Disord. 2013;14(1):339. doi:10.1186/1471-2474-14-339
35. Thompson DP, Urmston M, Oldham JA, Woby SR. The association between cognitive factors, pain and disability in patients with idiopathic chronic neck pain. Disabil Rehabil. 2010;32(21):1758-1767. doi:10.3109/09638281003734342
36. Walton DM, MacDermid JC, Giorgianni AA, Mascarenhas JC, West SC, Zammit CA. Risk Factors for Persistent Problems Following Acute Whiplash Injury: Update of a Systematic Review and Meta-analysis. J Orthop Sport Phys Ther. 2013;43(2):31-43. doi:10.2519/jospt.2013.4507
37. Martinez-Calderon J, Jensen MP, Morales-Asencio JM, Luque-Suarez A. Pain Catastrophizing and Function In Individuals With Chronic Musculoskeletal Pain. Clin J Pain. 2019;35(3):279-293. doi:10.1097/AJP.0000000000000676
38. Sullivan MJL, Bishop SR, Pivik J. The Pain Catastrophizing Scale: Development and validation. Psychol Assess. 1995;7(4):524-532. doi:10.1037/1040-3590.7.4.524
39. Osman A, Barrios FX, Gutierrez PM, Kopper BA, Merrifield T, Grittmann L. The pain catastrophizing scale: Further psychometric evaluation with adult samples. J Behav Med. 2000;23(4):351-365. doi:10.1023/A:1005548801037
40. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019;110:63-73. doi:10.1016/j.jclinepi.2019.02.016
41. Sauerbrei W, Perperoglou A, Schmid M, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagnostic Progn Res. 2020;4(1):3. doi:10.1186/s41512-020-00074-3
42. Kelly J, Ritchie C, Sterling M. Clinical prediction rules for prognosis and treatment prescription in neck pain: A systematic review. Musculoskelet Sci Pract. 2017;27:155-164. doi:10.1016/j.math.2016.10.066
43. Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HCW, 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. doi:10.1097/BRS.0b013e3181d85ad5
44. Wingbermühle RW, Heymans MW, van Trijffel E, Chiarotto A, Koes B, Verhagen AP. External validation of prognostic models for recovery in patients with neck pain. Brazilian J Phys Ther. 2021;25(6):775-784. doi:10.1016/j.bjpt.2021.06.001
45. Myhrvold BL, Kongsted A, Irgens P, Robinson HS, Thoresen M, Vøllestad NK. Broad External Validation and Update of a Prediction Model for Persistent Neck Pain after 12 Weeks. Spine (Phila Pa 1976). 2019;44(22):E1298-E1310. doi:10.1097/BRS.0000000000003144
46. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-247. doi:10.1016/j.jclinepi.2015.04.005
47. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):1-7. doi:10.1186/s12916-019-1466-7



Chapter 7
General discussion

7

Chapter 7. General discussion

Main findings of this thesis

The general aim of this thesis was to improve predictions of recovery of non-specific neck pain in individual patients in primary care with the use of prognostic prediction models. Two main research questions were evaluated to establish this aim. The first research question "Are valid prediction models available for making accurate predictions of recovery in patients with non-specific neck pain?" was addressed in **Chapter 2** and **Chapter 3**. Based on the evidence acquired from these chapters, the most common methodological challenges in prognostic modelling in spinal pain were reviewed and discussed in **Chapter 4**. The second research question: "Can newly developed prognostic models provide accurate predictions of recovery in primary care for patients with non-specific neck pain?" was subsequently addressed in **Chapter 5** and **Chapter 6**.

To evaluate if existing prediction models were able to make accurate predictions of recovery in people with non-specific neck pain, available prognostic models that could be used in primary care were first identified in a systematic review as described in **Chapter 2**. A large amount of 99 models for predicting recovery outcomes in people with neck pain was identified. These models were derived in 49 studies that all had a high risk of bias, especially relating to the participants' flow, analysis, and inappropriate sample size considerations. Reporting and methodological standards were often suboptimal with respect to model performance measures (e.g., calibration and discrimination), handling of missing data (e.g., multiple imputation), and dealing with overfitting (e.g., bootstrapping, shrinkage). In the vast majority of the studies, too many candidate predictors and categorical variables were considered during the modelling process relative to the number of events in the smallest outcome group, resulting in high risk of overfitting.

Seven models were subjected to external validation, four of them in high-risk of bias studies. Two Whiplash-Associated Disorder (WAD) models predicting disability [Neck Disability Index (NDI) at 6 and 12 months, respectively] and one non-specific neck pain model predicting perceived recovery at 6 months were evaluated at 6 or 12 months in low risk of bias studies and seemed promising for clinical use. However, the WAD models were evaluated in cohorts that did not solely contain primary care participants, and the non-specific neck pain model exhibited a limited discriminative ability (Area Under the Curve (AUC) of the receiver operating characteristic 0.65 (95% CI 0.59 to 0.71)) and a small pre-test to post-test probability shift. Therefore, it can be concluded that these three models needed further validation in a primary care setting before their clinical use could be advocated.

These models' predictive performance was subsequently evaluated in ANIMO, an external validation study using a large cohort of people with neck pain (n=1193) who were recruited and treated by manual therapists in Dutch primary care, as described in **Chapter 3**. Model performance in terms of discrimination and calibration appeared poor for the first WAD model and the non-specific neck pain model, with AUC's that were substantially below 0.70 and calibration slopes largely different from 1. The second WAD model could not be evaluated since several variables included in the model nor their proxies were available in the ANIMO dataset. The strength of this study was analysis in a large cohort by calibration and discrimination measures. Study limitations were the substantial number of missing values, and some variables that were not available in the validation set. It was concluded that external validation of these initially promising models was not successful and that their clinical use could not be recommended. This is an important message for clinicians since premature clinical use of an invalid prognostic model can lead to an inaccurate prognosis and subsequent suboptimal patient outcomes. These findings led us to the second research question whether newly developed prognostic models could provide accurate predictions of recovery in primary care for patients with non-specific neck pain.

During this research on prognostic models, besides common generic methodological shortcomings in the different healthcare fields, additional methodological challenges that are specific to the field of spinal pain were addressed. These were described as challenges with possible solutions in **Chapter 4**. The first challenge stated was the choice of participants, as differences in the selection of patients between studies may result in different case mix models that are difficult to compare and interpret. The second challenge was that, for prognostic modelling, data available from studies focussing on another aim is commonly used (e.g., randomized controlled trials), resulting in lacking potentially relevant predictor variables and variables that are not adequately operationalized. The third challenge concerned measurement limitations of patient-reported outcomes and predictors, which often are measured with patient-reported outcome measures (PROMs). A large variety of PROMs with different threshold cut-offs are being used in prognostic modelling studies. Measurement limitations such as insufficient construct validity, content validity, structural validity, and measurement error can influence PROMs' performance in a prognostic model. For example, various PROMs for disability that do not truly measure the same construct or use different threshold cut-offs to define outcomes, may result in models that include different predictors. Measurement error of patient-reported predictors is another example which can influence model performance and affect calibration and discrimination. The fourth challenge discussed was the complexity of the task of predicting recovery from spinal disorders. In fact, non-specific spinal disorders should typically be regarded as complex health problems with many interacting factors contributing to the course and prognosis of pain and disability. Current model-building approaches do not capture the still many unknown variables and their interactions involved, which also may change dynamically over time. The fifth and last challenge that was addressed was the confusion in prognostic factors of treatment response, which cannot be adequately evaluated in cohort study designs.

As there appeared to be no clinically useful prognostic model for recovery in people with neck pain and as updating these evaluated models was not useful considering their low predictive performance, we decided to develop and validate new prognostic models for recovery. This was done by emphatically preventing common methodological shortcomings which include reporting and handling of missing data, presentation of calibration and discrimination measures, and a priori sample size calculations. Furthermore, this implied

performing multiple imputation in case of missing data, preventing statistical selection of potential model predictors, and correcting for overfitting by bootstrapping and shrinkage. A model derivation study including internal validation was performed for the recovery of patients with neck pain, using the earlier described available large Dutch cohort data (n=1193), as we described in **Chapter 5**. Recovery was defined in terms of pain intensity, neck pain-related disability, and global perceived improvement immediately post-treatment and at 1-year follow-up. Discriminative performance was considered acceptable if AUC was ≥ 0.70. The developed post-treatment disability model exhibited the best predictive performance and potential for clinical use, showing a discriminative performance of AUC 0.74 (interquartile range (IQR), 0.72-0.75) after internal validation. Sensitivity analyses on complete cases showed comparable performance measure values. Most derived models yielded the same or almost the same predictors. The post-treatment disability model also contained sporting and previous episode predictors. The perceived improvement model contained fewer predictors. The developed post-treatment models for recovery of pain and perceived improvement initially reached acceptable performance, however, acceptable model performance thresholds were not achieved after internal validation.
None of the developed models for the prediction of recovery at 1-year reached acceptable performance and it is conceivable that long-term predictions are more difficult than short-term predictions. Predicting 6 weeks recovery has meaningful clinical value, since recovery from neck pain related disability at group level mainly takes place in the first 6 weeks without further subsequent improvement, and individuals indicated by the model as not recovered at 6 weeks may be at risk for long-term disability. [1] However, if a model was retrieved that predicts one-year recovery, this would have been of more clinical value since it provides the individual absolute risks for recovery at long-term.

Before clinical use of the post-treatment disability model can be recommended, the essential step of external validation was required.[2] Therefore, we conducted a broad external validation study in PRONEPA, a new Dutch cohort of people with (sub)acute neck pain (n=586) in primary care, treated with guideline-based usual care physiotherapy, as described in **Chapter 6**. Again, a discriminative performance of AUC ≥ 0.70 was considered acceptable. External validation of the recalibrated disability model at 6-week follow up was successful, with a discriminative performance of AUC 0.73 (95% CI: 0.69-0.77). At 12 weeks and post-treatment, it showed a nearly acceptable discriminative performance of AUC 0.69 (95% CI: 0.64-0.73) and 0.68 (95% CI: 0.63-0.72), respectively. The disability model was validated in a study that included people with neck pain for a maximum of 12 weeks. This limits the use of the model to people with acute and subacute neck pain, whereas the model cannot be used in people with chronic neck pain. Nonetheless, if using the model with subsequent intervention and advice leads to effective prevention of chronic neck pain, clinical implications will be important. This needs further evaluation of the model through impact analysis.

Additionally, it was of interest whether the derived models described in **Chapter 5** could be updated with physical examination and history variables that were available in the Dutch validation cohort, in particular cervical mobility, cervical anterior muscles endurance, and pain catastrophising. It appeared that cervical mobility added significant value to the disability model at all follow-up periods and pain catastrophising also to the 6-week pain model. However, these additional predictors improved model performance minimally and their measurement for purely prognostic purposes is burdensome considering the small gain in predictive performance.

In conclusion, after this broad external validation, it is suggested that physiotherapists use the disability model (without the additional predictors) at intake for the prognosis of people with neck pain to assist in clinical decisions concerning the recovery of neck pain-related disability at 6 weeks. Prognostic models can be presented as tools for clinical use e.g., through risk calculators, nomograms, or clinical prediction rules. [3] The disability model's regression formula is presented in Box 7.1, whereby predictions for clinical use are calculated by the linear predictor and successive logistic transformation into probabilities of persisting disability at 6 weeks. This can easily be transferred to a web-based risk calculator for clinical use, see https://www.somt.nl/research.
In this manner, the clinician is guided by the model to interpret the individual predicted risk to inform the patient about his/her prognosis to reach a shared decision. See Box 7.1. for an example in an individual patient.

Clinical scenario: a male patient born in 1971, presents for physiotherapy in primary care with neck pain complaints that are present since 21 days. After an initial screening, he would like to know his prognosis regarding persisting disability. He has problems sleeping, an NDI score of 23 and a FABQ-PA score of 11.

Validated Model formula (lp): -2.64 + 0.6 + (0.28*Subacute yes/no) + (0.11*Baseline NDI) + (0.02*Age) + (0.28*Sleeping problem yes/no) + (0.02*Baseline FABQ-PA)

Calculation formulas: ln (p/1-p) = 1/ (1+e^{-lp}); p = e^{lp}/1+ e^{lp}

Calculation:
Lp = -2.64 + 0.6 + (0.28*0) + (0.11*23) + (0.02*51) + (0.28*1) + (0.02*11) = -2.64 + 0.6 + 0 + 2.53 + 1.02 + 0.28 + 0.22 = 2.01
Ln (p/1-p) = 1/ (1+$e^{-2.01}$)
p = $e^{2.01}$/1+ $e^{2.01}$ = 7.463317/1+7.463317 = 0.88

Individual prognosis: probability of persisting disability of this patient at 6 weeks is 88%.

NDI = Neck Disability Index (scale 0-50)
FABQ-PA = Fear-Avoidance Believes Questionnaire, Physical Activity subscale (scale 0-24)
Lp = linear predictor
Ln = natural logarithm
P = individual's probability of persisting disability
e = Euler's number

Box 7.1. Illustration of using the validated disability model for individual risk predictions

In summary, this thesis provides insight into the fact that, despite a large number of published prognostic models for neck pain recovery (**Chapter 2**), the few positively evaluated models in earlier external validation studies appeared not valid in a successive external validation study in primary care (**Chapter 3**). Methodological shortcomings in prognostic modelling for patients with spinal disorders are highly common (**Chapter 4**). Therefore, currently available methodological standards in the new studies were closely followed. The model developed with up-to-date methodology predicting recovery of neck pain disability (**Chapter 5**) performed acceptably in a broad external validation study (**Chapter 6**). Physiotherapists are advised to use this model to counsel their patients about their likely short-term recovery regarding disability.

Limitations regarding this thesis

An important limitation in this thesis was the derivation of the disability model in data with a large amount of missing outcome (**Chapter 5**) that, despite the large sample size, may have put the derived disability model at risk of overfitting and may have some reliability implications. We believe to have accounted for this shortcoming by carefully applied multiple imputation, internal validation, and sensitivity analysis procedures. The sensitivity analysis may have revealed some reliability implications since the complete cases also contained the previous episode and sporting predictors. However, model evaluation in data with a very low amount of missing data and sufficient power (**Chapter 6**) showed acceptable model performance in broad external validation, which may indicate model robustness.
Furthermore, the results are limited in terms of the number of derived models that performed acceptably. Three post-treatment models and three 1-year models were developed and only the post-treatment disability model exhibited acceptable predictive performance. The other two post-treatment models (i.e., for recovery of pain and perceived improvement) lost their acceptable performance after internal validation (**Chapter 5**) and could not be updated to reach acceptable performance in the external validation study (**Chapter 6**). These two models contained one predictor parameter too many regarding the number of participants with the outcome in the development study which may lead to overfitting and may explain why their initially acceptable performance was not upheld after internal validation. Furthermore, the performance of these models may have been influenced more by the outcome measured with the Numeric Rating Scale (NRS) for pain and the General Perceived Effect (GPE) for perceived improvement, than NDI for disability. NDI is an instrument that may cover various health constructs. The NRS is a single-item questionnaire which measures a narrow aspect of the pain construct and may also have a larger measurement error. [4] The GPE was shown to reflect the current health status more than a change in health status over time. [5]

Making prognoses in clinical practice

Traditionally, the clinical reasoning process consists of consecutive clinical decisions on diagnosis, prognosis, and treatment. However, clinical judgement is susceptible to errors (e.g., reliability, measurement error) and prone to many cognitive biases (e.g., anchoring bias, conformation bias, availability bias, base rate neglect) and clinical decisions show

a considerable level of uncertainty and clinical variation. [6] [7] Error and bias may easily lead to false or at least uncertain predictions in prognostic reasoning and to subsequent inappropriate shared treatment decisions. As described in the introduction section of this thesis, the prognosis of non-specific neck pain is generally unfavourable, whereby 47% of the people with acute non-specific neck pain in general practice still report neck pain at 1-year follow-up. [8] People tend to underestimate the baseline risk (i.e., base rate neglect) and the amount of clinical uncertainty in prognosis is underestimated by clinicians, which can lead to distorted predictions. [7] In addition, clinicians and patients themselves tend to overestimate the beneficial effects of treatment and underestimate treatment risks. [9] [10]

A dominant model for decision-making regarded as applicable in clinical judgements is the dual process theory model. [11] The dual process theory model distinguishes a fast largely unconscious process in response to cues based on ingrained heuristics and a slow conscious reflective process of deliberative thinking. [12]
Deliberative thinking involving appropriate knowledge and cognitive debiasing may override the more error-sensitive fast heuristic processing [11] [12] However, cognitive debiasing is time-consuming and requires clinicians' awareness and willingness to change. [13] Using statistical models can be effective and efficiently to overcome errors in clinical judgement and is beneficial in the deliberative clinical reasoning thinking process. [14] [15] [16]

In people with neck pain presenting in primary care, a prognostic-oriented approach may be more useful compared to the classic diagnostic-oriented approach. [17] [18] In the vast majority of neck pain patients, their pain is no symptom of serious pathology that needs an immediate intervention to avoid a serious outcome (and prevent a worse prognosis) and it is labelled after initial screening assessment, as non-specific neck pain. This limits the information to the presence of a global pain location only. Also, the commonly used diagnostic classification system proposed by the Neck Pain Task Force (NPTF) provides limited information. [19] If a patient is classified by this system as neck pain grade 2 (no signs or symptoms of major structural pathology, but major interference with activities of daily living), [19] information on the extent of interference and type of activities is lacking. Furthermore, valuable prognostic factors for recovery and factors predicting the beneficial effect of (specific) treatment are missing in this classification system and could be overlooked in case a clinician's focus is mainly on diagnosis. In its essence, the aim of the clinical reasoning process in non-specific neck pain is to optimise the patient's individual prognostic outcome(s), by addressing (e.g., treatment, providing advice) all personal, treatment, and contextual factors that affect the patient's specific outcome(s). These may concern patient-specific biopsychosocial factors from e.g., history taking, questionnaires, physical examination, and factors predicting the effect of treatment. Predicting recovery of non-specific neck pain involving these many interacting factors is complex, as we described in **Chapter 4**. Even with deliberative thinking to reduce biases, clinicians' information processing is limited, and statistical models are capable of processing more information simultaneously. [14] Prognostic and predictive models can be more accurate than predictions based on clinical judgement alone. [14] [15] [16] Integrating statistical models and clinical judgement has been advised for reliable prognosis and prediction.[20] [13]

Clinical practice guidelines intend to reduce practice variation, error, and biases. The inclusion of accurate prognostic and predictive models in clinical reasoning and clinical guidelines may further reduce this variation. Guidelines on neck pain stimulate the use of prognostic factors but rarely provide information on prognostic models or provide clear direction for their clinical use. [21] The recently updated guidelines on neck pain published by the American Physical Therapy Association provide some information on prognostic models for recovery and prediction models for treatment effect. [22 23]These guidelines stress that validation is required before widespread clinical use of prognostic models can be recommended. [22] The current guideline on neck pain of the Royal Dutch Society for Physical Therapy recommends analysing modifiable prognostic factors that can explain a deviant course of complaints. [24] The existence of prognostic models is mentioned in the guideline and the score chart of the model published by Schellingerhout et. al [25] is presented in an accompanying explanation file. However, the guideline provides no advice on its use. [24] We recommend against its use in primary care, since Schellingerhout's model showed poor performance in **Chapter 3** and more recently in another study.[26] Generally, a model can be cautiously considered for clinical use after successful external validation [27] after following methodological standards for designing, executing, and reporting as described in **Chapter 4.** However, accurate model performance does not guarantee clinical utility. [3 28 29] Therefore, it is advisable to evaluate a model's clinical utility (e.g., with decision curve analysis) before its inclusion in clinical guidelines.

Individual prognostic factors still provide information on the prognosis of people with neck pain in general, while an accurate prognostic model is preferable since it has the advantage of providing an individual prognosis for a specific patient with neck pain consulting a clinician. As stated in the general introduction section (**Chapter 1**) of this thesis, a prognostic model can be considered for use in clinical practice after successful external validation in a setting comparable to its intended use. [27 29]

Since the developed disability model described in **Chapter 5** appeared sufficiently accurate in a broad external validation study (**Chapter 6**), its use can be recommended in a primary care physiotherapy setting to counsel individual patients with acute and subacute non-specific neck pain about their short-term prognosis of recovery regarding disability. Caution is still necessary, however, as the model's clinical utility has not been evaluated. Therefore, using the disability model is recommended to assist in shared clinical decisions and not to establish clinical decisions based purely on the model.

Currently, the clinical use of any other prognostic model cannot be recommended since no other accurate valid prognostic model is available (**Chapter 2 and Chapter 3**). The disability model can be used to indicate a person with neck pain that has a high or a low probability of recovery from neck pain related disability at 6 weeks, with reasonable accuracy (see Box 7.1). A physiotherapist and patient can use this prognostic information as input for shared decisions on short-term recovery already during intake and this may support personalized treatment considerations. For instance, in case of a high chance of recovery of disability at 6 weeks in a patient with acute or subacute neck pain, the clinician (e.g., general practitioner, physiotherapist) and patient in dialogue may consider a reassuring and self-management approach, instead of providing treatment. However, in case of a very low chance of recovery of disability in the same patient, the clinician may timely shift to a focus on multi-modal treatment options and discuss these with the patient, facilitated using the model. Elaboration of treatment efficacy is beyond the scope of this thesis, but interventions may include, besides exercise therapy and manual therapy [30 31 32], various recommended interventions such as pain neuroscience education, graded exposure therapy or cognitive behavioural therapy targeting the identified personal psychosocial factors. [33 34]

The validated model described in **Chapter 5** predicts neck pain related disability using the NDI outcome measurement which is believed to reflect both the constructs of pain and disability. [35 36] Patients can have different perspectives on the construct of recovery which is critical to the used outcome measurement, and it follows that the model is not directly suitable for patients' pursuing recovery for e.g., return to work, specific functions or activities, or quality of life. [37] A related perspective is to what extent does the patient wish to recover exactly? The patient can expect to fully recover, or the patient and the clinician may decide that a certain amount of improvement is a more realistic future goal. The disability model's outcome criterion for recovery was cut-off at NDI < 8% in the development and validation studies, which means that the model is too strict for a patient who is satisfied with a recovery amount up to a value of 20%. Also, the model is not suitable for a certain amount of improvement (e.g., improving from NDI 40 to 20). In conclusion, it is important that the clinician appraises which perspective on recovery the specific individual patient has and ascertain this accommodates the model's outcome criterion for recovery.

Recommendations for research

Some researchers have indicated that statistical models in general are superior to predictions based on clinical judgement. [14 15 16] However, this has not been evaluated systematically in medicine or any musculoskeletal domain. New studies comparing the performance of clinical judgments on prognosis in non-specific neck pain against the performance of statistical models, or against combining both, could provide useful information for guideline development and recommendations. Sufficient statistical model performance of the disability model at 6 weeks does not guarantee that the model is useful in clinical situations. [3 28 29] Additionally, a moderately performing model can be useful if it provides accurate predictions across certain risk thresholds. The disability model at 12-weeks showed a discriminative performance just below AUC 0.70 and was well calibrated (**Chapter 6**). Therefore, further assessment of the disability model is recommended at both 6 and 12 weeks with decision curve analysis for its ability to inform clinical decisions. [38 39 40] A decision curve weighs the benefit of treatment to harm by overtreatment and summarises model performance over a range of decision thresholds of predicted risk; also, it can compare if a specific clinical action considered from a decision based on concerning prognostic model is more beneficial than a 'treat all' or 'treat none' strategy. [41]
As mentioned in the general introduction section (**Chapter 1**) of this thesis, the ultimate step in model development is impact analysis. Finally, impact analysis studies are needed to evaluate if the use of the prognostic model for post-treatment disability can improve clinical decisions leading to better patients' outcomes and/or reducing costs. [3 28]

The participants in the derivation and validation studies described in **Chapters 3, 5 and 6**, received usual care treatment with manual therapy in the ANIMO study and physiotherapy in the PRONEPA study. Effective treatment reduces the risk of the undesired outcome. However, the effectiveness of treatment can vary between individuals, and this may have affected recovered and non-recovered groups unequally. Therefore, it is essential to identify predictors of treatment effect for individuals to make optimal treatment decisions. [28] Predictors of treatment effect are factors associated with the response to specific intervention and require different study designs (i.e., randomized clinical trials) as was described in **Chapter 4**. Predictors of treatment effect can be evaluated by investigating the interaction of that predictor with targeted intervention components as an additional effect on the outcome over and above that of the predictor and treatment alone. [28] For example, studying the interaction of cervical mobility with mobilisation or manipulation interventions, pain catastrophising with a cognitive-behavioural intervention, and fear-avoidance behaviour with a graded activity intervention in a cohort study as suggested in **Chapter 6** can provide information on potential predictors for treatment effect. Further evaluation should involve double-arm randomized controlled trials.

The disability model's relatively low explained variation of 20% indicates there are still predictors missing. Clinicians inform patients of their individual prognosis and experienced clinicians may retain important clinical knowledge on factors for predicting neck pain recovery from a clinical point of view. As mentioned earlier in this chapter, clinicians reach their clinical and prognostic judgements through the interplay between a fast intuitive and a slow more controlled way of thinking. [42] Qualitative research designed with the interaction between experienced clinicians with various backgrounds treating people with neck pain and stimulating a deliberate reflection on their prognostic reasoning process can provide valuable insight into clinical prognostic thinking and may reveal new prognostic factors for neck pain recovery. Also, patients may retain valuable information and qualitative research through interviews and focus groups with patients may reveal some new insights into prognostic factors from their perspective as well. Furthermore, it is of interest to know through qualitative research if patients value 6- and 12-weeks prognosis of neck pain related disability as important compared to long term prognosis, since the disability model predicts short term prognosis sufficiently accurate and there are no models available for long-term prognosis.

Furthermore, clinicians usually measure baseline factors at intake. Similarly, the models retrieved and validated in this thesis are suitable for prognostic judgments at intake. In research settings, however, factors regularly are gathered at several time points between intake and outtake. This allows for incorporating change scores in the models. No clinically changed disability status and no changed pain intensity scores were predictors retrieved in a good performing model, predicting low back pain at 3 and 6 months follow up. [43] Incorporating change score predictors may improve prognostic and predictive model performance in non-specific neck pain, which could be used at outtake for long-term prognosis.

Can we really predict the future of non-specific neck pain?

A prognostic model for non-specific neck pain was developed and validated, thereby emphatically following currently available methodological standards for model designing, executing, and reporting (**Chapters 5 and 6**). However, the results are limited in terms of the number of derived models, their relatively low explained variation, and the fact that only for one outcome (disability) the short-term models reached acceptable performance.

This raises the question if following the recommendations in this thesis for further research and pursuing the path geared to emphatically following methodological standards, will lead to increased disability model accuracy, valid prognostic models for other outcomes, and long-term prognosis. However, there are examples of prognostic models in the medical field that have passed many validation studies and are cited in national guidelines indicating their use e.g., the Nottingham Prognostic Index for breast cancer survival, and the GRACE score for mortality/myocardial infarction after acute coronary syndrome. [28] It is of notice, that prognostic models in medicine are developed within a medical framework of disease classification, mostly resulting in the use of objective outcome measurements defined in terms of mortality and disease. In contrast, the physiotherapy field classifies health problems within a framework of the international classification of Function, Disability and Health [44] with many interacting factors and prevailing use of subjective outcome (and predictor) measurements, which challenges long-term recovery predictions. Nevertheless, we developed and validated a reasonably performing model using a dataset with several limitations. Other researchers retrieved good-performing models for predicting non-recovery of back pain at 6 months and 1-year follow-up. [45] Therefore, using a more comprehensive dataset of high quality, especially gathered for prognostic modelling, has the potential for better performance of prognostic models for recovery of non-specific neck pain at medium and long-term follow-up.

Additionally, large studies with high-quality data, analysed with artificial intelligence and machine learning techniques may provide further solutions for this challenge. However, machine learning techniques are 'data hungry', and many have a black box nature, raising methodological concerns e.g., overfitting, lack of validation, and lack of transparency of the computer algorithm.[46] [47] This currently limits the generalizability and usability of the study findings and there is a need for specific methodological standards. In response, an extension of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement and Prediction Model Risk Of Bias Assessment Tool (PROBAST) for reporting and critical appraisal of prognostic model studies for machine learning techniques is now being developed. [48] If these and future methodological standards for designing, executing, and reporting machine learning techniques are followed, the very large quantities of data -that will become increasingly available by routinely collected data in electronic healthcare records, wearables, and social apps- may be captured with artificial intelligence for prognostic purposes. At the same time, routinely collected data in primary care is not necessarily of high quality. For example, in Dutch physiotherapy, electronic healthcare record data including PROMs e.g., NRS, NDI are routinely collected nationally as contracted with healthcare insurance companies and quality networks.

Currently, this data is used for providing feedback to clinicians and not for research purposes, which may be challenging. Quality control and processing of current routinely collected electronic healthcare data are now unclear, and the essential data quality may not be at the desired level. Therefore, data quality should be controlled prior to using this data for research purposes which may require a general culture shift in primary care clinical physiotherapy.

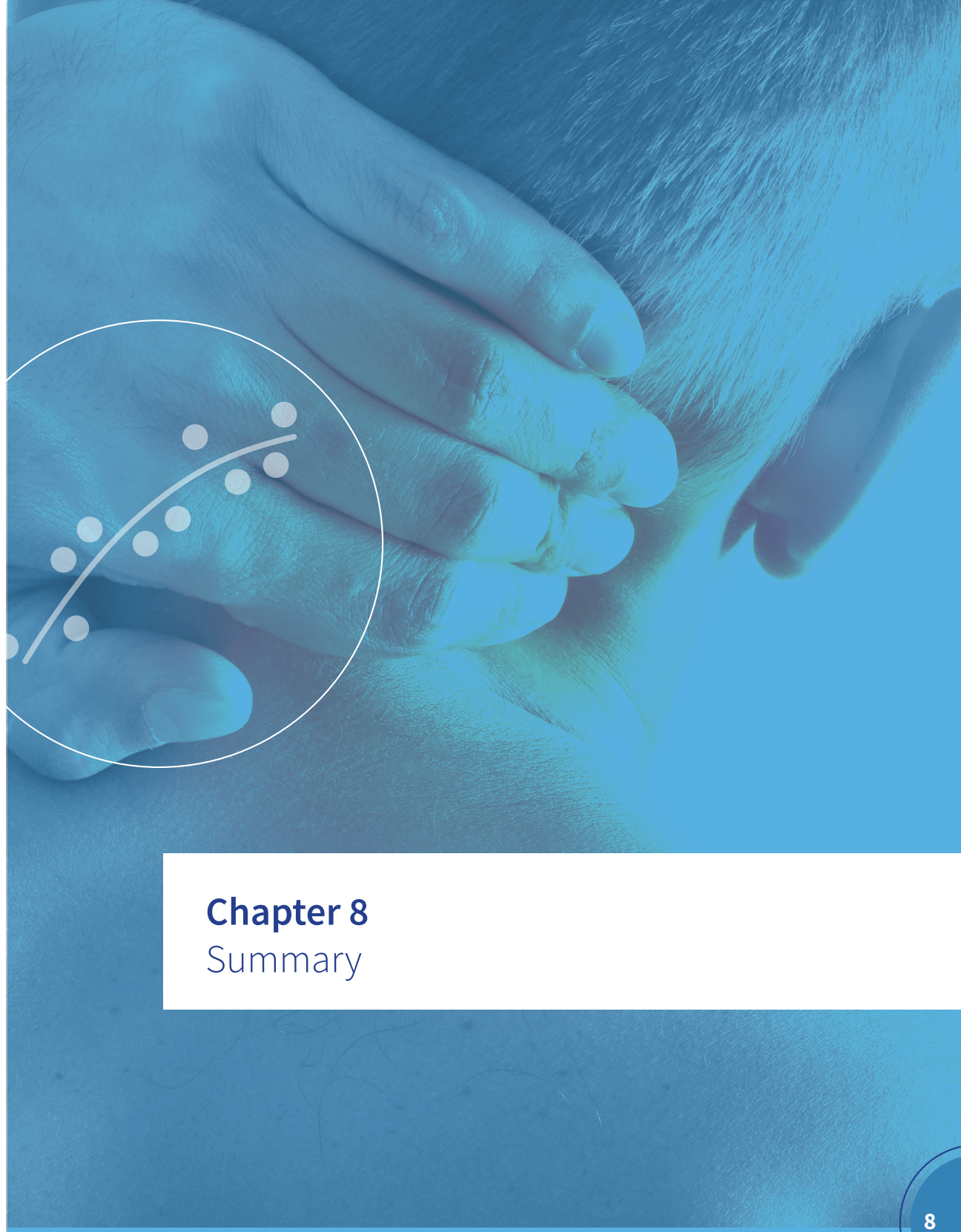
Conclusions

The aim of this thesis to improve predictions of recovery of non-specific neck pain in individual patients in primary care with the use of prognostic prediction models was fairly reached.
There appeared no valid prognostic models available, and we decided to develop and validate a new prognostic model. The developed model predicting short-term recovery of disability in people with non-specific neck pain performed well at broad external validation. Using this prognostic model in a primary care setting can reduce practice variation and it is recommended to assist clinical decisions and counsel individual patients with acute and subacute non-specific neck pain about their short-term prognosis of recovery regarding disability. Reflecting on the clinical challenge that inspired me to conduct this thesis, including this model in my clinical reasoning reduces some uncertainty of short-term prognosis regarding disability and provides direction to answer my patient's prognostic questions. Further research can help answer remaining questions regarding the prognosis of long-term disability and other outcomes and the clinical impact of the developed model.

References

1. Hush JM, Lin CC, Michaleff Z a, Verhagen A, Refshauge KM. Prognosis of Acute Idiopathic Neck Pain is Poor: A Systematic Review and Meta-Analysis. Archives of Physical Medicine and Rehabilitation. 2011;92(5):824-829. doi:10.1016/j.apmr.2010.12.025
2. Collins GS, de Groot J a, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14(1):40. doi:10.1186/1471-2288-14-40
3. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer Science and Business Media,; 2019.
4. 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. Journal of Pain. 2019;20(3):245-263. doi:10.1016/j.jpain.2018.07.009
5. 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. Journal of Clinical Epidemiology. 2010;63(7):760-766.e1. doi:10.1016/j.jclinepi.2009.09.009
6. Bell I, Mellor D. Clinical judgements: Research and practice. Australian Psychologist. 2009;44(2):112-121. doi:10.1080/00050060802550023
7. Makridakis S, Kirkham R, Wakefield A, Papadaki M, Kirkham J, Long L. Forecasting, uncertainty and risk; perspectives on clinical decision-making in preventive and curative medicine. International Journal of Forecasting. 2019;35(2):659-666. doi:10.1016/j.ijforecast.2017.11.003
8. Vos CJ, Verhagen AP, Passchier J, Koes BW. Clinical course and prognostic factors in acute neck pain: an inception cohort study in general practice. Pain Medicine. 2008;9(5):572-580 9p. doi:10.1111/j.1526-4637.2008.00456.x
9. Hanoch Y, Rolison J, Freund AM. Reaping the Benefits and Avoiding the Risks: Unrealistic Optimism in the Health Domain. Risk Analysis. 2019;39(4):792-804. doi:10.1111/risa.13204
10. Hoffmann TC, del Mar C. Clinicians' Expectations of the Benefits and Harms of Treatments, Screening, and Tests. JAMA Internal Medicine. 2017;177(3):407. doi:10.1001/jamainternmed.2016.8254
11. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origind of bias and theory of debiasing. BMJ Quality & Safety. 2013;22(10):58-64. doi:10.1136/bmjqs-2013-002387
12. Kahneman D. Thinking, Fast and Slow. Farrar, Straus and Giroux; 2011.
13. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 2: impediments to and strategies for change. BMJ Quality & Safety. 2013;22(10):789-792. doi:10.1136/bmjqs-2013-002387
14. van Bronswijk SC, Lemmens LHJM, Huibers MJH, Peeters FPML. Selecting the optimal treatment for a depressed individual: Clinical judgment or statistical prediction? Journal of Affective Disorders. 2021;279:149-157. doi:10.1016/j.jad.2020.09.135
15. Ægisdóttir S, White MJ, Spengler PM, et al. The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. The Counseling Psychologist. 2006;34(3):341-382. doi:10.1177/0011000005285875
16. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment. 2000;12(1):19-30. doi:10.1037/1040-3590.12.1.19
17. Croft P, Dinant GJ, Coventry P, Barraclough K. Looking to the future: should 'prognosis' be heard as often as 'diagnosis' in medical education? Education for Primary Care. 2015;26(6):367-371. doi:10.1080/14739879.2015.1101863
18. Aasdahl L, Granviken F, Meisingset I, Woodhouse A, Evensen KAI, Vasseljen O. Recovery trajectories in common musculoskeletal complaints by diagnosis contra prognostic phenotypes. BMC Musculoskeletal Disorders. 2021;22(1):455. doi:10.1186/s12891-021-04332-3
19. Guzman J, Hurwitz EL, Carroll LJ, 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. Journal of Manipulative and Physiological Therapeutics. 2009;32(2 SUPPL.):17-28. doi:10.1016/j.jmpt.2008.11.007
20. Soyiri IN, Reidpath DD. An overview of health forecasting. Environmental Health and Preventive Medicine. 2013;18(1):1-9. doi:10.1007/s12199-012-0294-6
21. Parikh P, Santaguida P, Macdermid J, Gross A, Eshtiaghi A. Comparison of CPG's for the diagnosis, prognosis and management of non-specific neck pain: a systematic review. BMC Musculoskeletal Disorders. 2019;20(1):81. doi:10.1186/s12891-019-2441-3
22. Blanpied PR, Gross AR, Elliott JM, et al. Clinical practice guidelines linked to the international classification of functioning, disability and health from the orthopaedic section of the American physical therapy association. Journal of Orthopaedic and Sports Physical Therapy. 2017;47(7):A1-A83. doi:10.2519/jospt.2017.0302

23. Childs JD, Cleland J a, Elliott JM, et al. Neck pain: Clinical practice guidelines linked to the International Classification of Functioning, Disability, and Health from the Orthopedic Section of the American Physical Therapy Association. J Orthop Sports Phys Ther. 2008;38(9):A1-A34. doi:10.2519/jospt.2008.0303
24. Bier JD, Scholten-Peeters GGM, Staal JB, et al. KNGF-richtlijn nekpijn, verantwoording en toelichting. https://www.kngf.nl/kennisplatform/richtlijnen/nekpijn
25. Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HCW, 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. doi:10.1097/BRS.0b013e3181d85ad5
26. Myhrvold BL, Kongsted A, Irgens P, Robinson HS, Thoresen M, Vøllestad NK. Broad External Validation and Update of a Prediction Model for Persistent Neck Pain after 12 Weeks. Spine (Phila Pa 1976). 2019;44(22):E1298-E1310. doi:10.1097/BRS.0000000000003144
27. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: A review. Journal of Clinical Epidemiology. 2008;61(11):1085-1094. doi:10.1016/j.jclinepi.2008.04.008
28. Riley RD, van der Windt DA, Croft P, Moons KGM. Prognosis Research in Health Care, Concepts, Methods, and Impact. 1st ed. Oxford University Press; 2019.
29. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Medicine. 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381
30. 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(6):e0178621. doi:10.1371/journal.pone.0178621
31. Coulter ID, Crawford C, Vernon H, 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(2):E55-E70.
32. Gross A, Langevin P, Burnie SJ, et al. Manipulation and mobilisation for neck pain contrasted against an inactive control or another active treatment. Cochrane Database of Systematic Reviews. 2015;(October). doi:10.1002/14651858.CD004249.pub4
33. Watson JA, Ryan CG, Cooper L, et al. Pain Neuroscience Education for Adults With Chronic Musculoskeletal Pain: A Mixed-Methods Systematic Review and Meta-Analysis. The Journal of Pain. 2019;20(10):1140.e1-1140.e22. doi:10.1016/j.jpain.2019.02.011
34. Monticone M, Cedraschi C, Rocca B, et al. Cognitive-behavioural treatment for subacute and chronic neck pain. In: Monticone M, ed. Cochrane Database of Systematic Reviews. Vol 2013. John Wiley & Sons, Ltd; 2013. doi:10.1002/14651858.CD010664
35. MacDermid JC, Walton DM, Avery S, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther. 2009;39(5):400-417. doi:10.2519/jospt.2009.2930
36. 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. Journal of Clinical Epidemiology. 2013;66(7):775-782.e2. doi:10.1016/j.jclinepi.2013.02.005
37. 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 & Research. 2008;61(1):124-131. doi:10.1002/art.24162
38. van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. European Urology. 2018;74(6):796-804. doi:10.1016/j.eururo.2018.08.038
39. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research. 2019;3(1):18. doi:10.1186/s41512-019-0064-7
40. van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. European Urology. 2018;74(6):796-804. doi:10.1016/j.eururo.2018.08.038
41. Traeger AC, Hübscher M, McAuley JH. Understanding the usefulness of prognostic models in clinical decision-making. Journal of Physiotherapy. 2017;63(2):121-125. doi:10.1016/j.jphys.2017.01.003
42. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. BMJ Quality & Safety. 2013;22(Suppl 2):ii58-ii64. doi:10.1136/bmjqs-2012-001712
43. Heymans MW, van Buuren S, Knol DL, Anema JR, van Mechelen W, de Vet HCW. The prognosis of chronic low back pain is determined by changes in pain and disability in the initial period. The Spine Journal. 2010;10(10):847-856. doi:10.1016/j.spinee.2010.06.005
44. World Health Organisation (WHO). International classification of functioning, disability and health (ICF). https://www.who.int/classifications/international-classification-of-functioning-disability-and-health. Accessed June 25, 2022. https://www.who.int/classifications/international-classification-of-functioning-disability-and-health
45. van der Gaag WH, Chiarotto A, Heymans MW, et al. Developing clinical prediction models for nonrecovery in older patients seeking care for back pain: the back complaints in the elders prospective cohort study. Pain. 2021;162(6):1632-1640. doi:10.1097/j.pain.0000000000002161
46. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. The Lancet. 2019;393(10181):1577-1579. doi:10.1016/S0140-6736(19)30037-6
47. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. New England Journal of Medicine. 2017;376(26):2507-2509. doi:10.1056/nejmp1702071
48. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008. doi:10.1136/bmjopen-2020-048008



Chapter 8
Summary

Chapter 8. Summary

Prognosis of neck pain

Neck pain is a common global health problem leading to substantial pain, disability, and economic costs in most countries. This also applies to neck pain in primary care in the Netherlands, which shows high incidence and prevalence numbers in general practice and is the second most registered diagnostic code by physiotherapists. Neck pain is usually divided into specific and non-specific neck pain. The vast majority of neck pain concerns conditions without an identifiable pathoanatomical cause and are thus labelled as nonspecific. The prognosis of non-specific neck pain is after a few weeks generally unfavourable. Recovery of neck pain mainly takes place in the first 4-6 weeks, without further evident reduction of neck pain and disability afterwards. In general practice in the Netherlands, 47% of acute non-specific neck pain patients reported still having neck pain at 1-year follow-up. This indicates that, when people do not recover within the first few weeks, prognosis leads for a substantial proportion of people to persistent or intermittent pain and disability. Identification of patients very likely to recover in the short term may reduce the risk of overtreatment and health costs. Moreover, early identification of neck pain patients with expected worse outcomes enables clinicians to offer effective treatments timely and may abate patient's burden and health costs.
Patients with neck pain have concerns about their future and like to know their prognosis when consulting their primary care clinician. The prognosis of recovery of non-specific neck pain in individual patients is a challenging task for a clinician. Prognostic factors and prognostic models can provide a clinician with additional information to improve the estimation of the patients' individual prognosis. Prognostic factors yet provide information on the prognosis of people with neck pain in general, while an accurate prognostic model is preferable since it has the advantage of providing an individual prognosis for a specific patient. Prognostic models are used for providing an individual prognosis by clinicians in various healthcare domains and settings and could be useful for the prognosis of non-specific neck pain in primary care. 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. The shift to personalized medicine has led to a vast amount of published prognostic models.

The general aim of this thesis was to improve predictions of recovery of non-specific neck pain in individual patients in primary care with the use of prognostic prediction models.
Chapter 1 is an introduction to the research conducted in this thesis and the two main research questions that were evaluated to establish this aim are described.
The two research questions were:

1. Are valid prediction models available for making accurate predictions of recovery in patients with non-specific neck pain? (Addressed in **Chapter 2 and Chapter 3)**. Based on the evidence acquired from these chapters, the most common methodological and additional challenges in prognostic modelling in spinal pain were reviewed and discussed in **Chapter 4.**
2. Can newly developed prognostic models provide accurate predictions of recovery in primary care for patients with non-specific neck pain? (Addressed in **Chapter 5 and Chapter 6)**.

In **Chapter 2** a systematic review was performed to summarize existing multivariable prognostic models for recovery in people with non-specific neck pain that could be used in primary care. Studies were included when the outcome concerned pain reduction, reduced disability, or perceived recovery at any time of follow-up. Fifty-three publications were included, of which 46 were derivation studies, four validation studies, and three combined studies. We evaluated the quality of the selected studies using the novel Prediction model study Risk of Bias Assessment Tool (PROBAST). The PROBAST was designed to assess the risk of bias, applicability, and usability of multivariable prediction model studies included in a systematic review.
A large amount of 99 models for predicting recovery outcomes in people with neck pain was identified. These models were derived in 49 studies that all had a high risk of bias. This was especially related to the participant's flow, analysis, and inappropriate sample size considerations.
Reporting and methodological standards were often suboptimal with respect to model performance measures, handling of missing data, and dealing with overfitting.
Seven models were subjected to external validation, four of them in high risk of bias studies. Three externally validated models that were evaluated at 6 or 12 months follow-up, generated models in low risk of bias studies: Two Whiplash-Associated Disorder (WAD) models predicting disability (Neck Disability Index (NDI) at 6 and 12 months follow-up, respectively) and one non-specific neck pain model predicting perceived recovery at 6 months follow-up. These three models seemed promising for clinical use. However, the WAD models were evaluated in cohorts that did not solely contain primary care participants, and the non-specific neck pain model exhibited a limited discriminative ability (Area Under the Curve (AUC) of the receiver operating characteristic 0.65 (95% CI 0.59 to 0.71)). Therefore, we concluded that these three models needed further validation in a primary care setting before their clinical use could be advocated.

In **Chapter 3** these three promising models' predictive performance in terms of discrimination and calibration was evaluated in an external validation study. The findings of this study were reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) recommendations. Data from the ANIMO study was used. ANIMO is a prospective cohort study that ran from October 2007 until March 2008, that aimed to describe usual care manual therapy for patients with neck pain in the Netherlands and explored outcomes and adverse events of treatment. It concerned a large cohort of people with neck pain (n=1193) who were recruited and treated by directly assessable manual therapists in Dutch primary care.
Received treatment consisted of usual care manual therapy and may have included specific joint mobilizations, high-velocity thrust techniques, myofascial techniques, giving advice, or specific exercises.
Model performance in terms of discrimination and calibration appeared poor for one WAD model and the non-specific neck pain model, with AUC's that were substantially below 0.70

and calibration slopes largely different from 1. The other WAD model could not be evaluated since several variables included in the model nor their proxies were available in the ANIMO dataset. We concluded that external validation of these initially promising models was not successful and that their clinical use could not be recommended.

In **Chapter 4** methodological shortcomings in prognostic model research are described. During our research on prognostic models, common generic methodological shortcomings in the field of healthcare appeared highly common in the field of spinal pain. Common generic methodological shortcomings also identified in the spinal field are too many candidate predictors categories relative to the number of events, predictor selection based purely on statistical significance, categorization of continuous predictors, lack of reporting of key performance measures and poor overall reporting. These common shortcomings often lead to overfitted, over-optimistic or unstable models. This results in models that generalize poorly to other clinical settings and patients. These common shortcomings can to a large extent be addressed by following currently available methodological standards for designing, executing, and reporting prediction models in healthcare.
Furthermore, additional methodological challenges that are specific to the field of spinal pain were encountered. Based on the evidence acquired during this research, five additional methodological challenges in prognostic modelling in spinal pain were reviewed and discussed. The first challenge was related to the choice of participants. Differences in the selection of patients between studies may result in different case mix models that are difficult to compare and to interpret. To counter this problem, there should be a clear description of recruitment and selection criteria, with a demarcation of subgroups with expected different prognoses (e.g., WAD Grade III). The second challenge was related to the purpose of the studies. For prognostic modelling, data available from studies focussing on another aim is commonly used. This often results in lacking potentially relevant predictor variables and variables that are not adequately operationalized. This problem can be countered by using large data sets purposively designed for prognostic modelling which contain a large array of potentially relevant biological, physical, and psychosocial candidate predictors. The third challenge was related to limitations in the measurement of outcomes and predictors.

These often are measured with patient-reported outcome measures (PROMs) and a large variety of PROMs with different threshold cut-offs are being used. The development of consensus-based core outcome sets for prognostic models in spinal pain may be a solution to deal with this heterogeneity. Measurement limitations such as insufficient construct validity, content validity and structural validity, can influence PROMs' performance in a prognostic model. Furthermore, measurement error of self-reported predictors appears to influence model performance; random error decreases calibration and discrimination, whereas systematic error affects calibration and does not influence discrimination. The fourth challenge was related to the complexity of recovery predictions. Non-specific spinal disorders should typically be regarded as complex health problems with many interacting factors contributing to the prognosis of pain and disability. Current model-building approaches do not capture the still many unknown variables and their interactions, which also may change dynamically over time. Including interaction and predictor trajectory over time variables during model-building, has the potential to improve model performance. We envisioned that artificial intelligence and machine learning techniques will be capable of discovering and modelling prognostic factors and their interactions in large data sets, linking data from various recourses. The fifth challenge was related to the confusion of prognosis and treatment response. Prognostic factors do not necessarily also predict the effect of treatment. Predictors of treatment effect are evaluated by investigating the interaction of that predictor with treatment as an additional effect on the outcome. Single-arm cohort studies may provide exploratory information. However, double-arm trials are needed for stronger model development and validation.

The findings of the poor model performance in the external validation study (**Chapter 3)**, led us to the second research question whether newly developed prognostic models provide accurate predictions of recovery in primary care for patients with non-specific neck pain. We decided to develop and validate new prognostic models for recovery, by emphatically preventing common methodological shortcomings.
The objective of **Chapter 5** was to develop and internally validate models for recovery of patients with non-specific neck pain of any duration. We used the earlier described available large ANIMO cohort data that consisted of adults (n=1193) recruited and treated by manual therapists in Dutch primary care. The outcome measures used to define recovery were pain intensity, neck pain-related disability, and global perceived improvement immediately post-treatment and at 1-year follow-up. Fourteen to eighteen candidate predictor categories were considered in the multivariable analyses for the six models. Discriminative performance was considered acceptable if AUC was ≥ 0.70. The 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; interquartile range (IQR), 0.91–0.93). The model showed a discriminative performance of AUC 0.74 (IQR: 0.72-0.75) after internal validation and has the best potential for clinical use. The initially reached acceptable performances of the post-treatment models for recovery of pain and perceived improvement were not achieved after internal validation. None of the developed models for prediction of recovery at 1-year reached acceptable performance.

In **Chapter 6** we conducted a broad external validation study of the prognostic models in the PRONEPA cohort, that were developed in **Chapter 5**. The models were evaluated at post-treatment, 6 and 12 weeks follow-up. PRONEPA is a prospective cohort study that ran from November 2020 until April 2021 in the Netherlands, that primarily aimed to evaluate prognostic factors that predict the development of chronic neck pain in people with (sub) acute neck pain (n=586). Participants were registered for primary care physiotherapy and recruited and treated by directly assessable physiotherapists who were graduating from a Master of Science program in manual therapy. Received treatment consisted of guideline-based usual care physiotherapy. Discriminative performance was considered acceptable if AUC was ≥ 0.70. External validation of the disability model at 6 weeks showed a discriminative performance of AUC 0.73 (95% CI: 0.69-0.77) and a reasonably well calibration after intercept recalibration. External validation of the disability model at 12 weeks and at post-treatment showed nearly acceptable discriminative performance of AUC 0.69 (95% CI: 0.64-0.73) and 0.68 (95% CI: 0.63-0.72), respectively, and was well calibrated.

Additionally, it was of interest whether the derived models described in **Chapter 5** could be updated with variables that were available in the dataset of the PRONEPA cohort, in particular cervical mobility, cervical anterior muscles endurance, and pain catastrophising. It appeared that cervical mobility added significant value to the disability model at all follow-up periods and pain catastrophising also to the 6-week pain model. However, these additional predictors improved model performance minimally.
We suggested that physiotherapists use the disability model, without the additional predictors, at intake for the prognosis of people with neck pain to assist in clinical decisions concerning the recovery of neck pain-related disability at 6 weeks. Further research is needed to assess the disability model's clinical impact.

Finally, in **Chapter 7**, we reflected on the main findings in this thesis and making a prognosis in clinical practice and elaborated on implications for clinical practice and research.

Samenvatting

Prognose van nekpijn

Nekpijn is een wereldwijd gezondheidsprobleem wat in de meeste landen leidt tot aanzienlijke pijn, beperkingen in het dagelijks functioneren plus economische kosten. Dit geldt ook voor nekpijn in de eerstelijnsgezondheidszorg in Nederland, die bij huisartsen een hoge incidentie en prevalentie kent en de op een na meest geregistreerde code is bij fysiotherapeuten. Nekpijn wordt meestal onderverdeeld in specifieke- en niet-specifieke nekpijn. In het overgrote deel van de mensen met nekpijn is geen pathologisch-anatomische oorzaak aantoonbaar en wordt het gelabeld als niet-specifiek. De prognose van nekpijn is, als de klachten niet binnen een paar weken herstellen, over het algemeen ongunstig. Herstel van nekpijn treedt vooral op in de eerste vier tot zes weken, daarna is er geen evidente afname meer van pijn en beperkingen in het dagelijks functioneren. De helft van de patiënten met niet-specifieke nekpijn in de Nederlandse huisartsenpraktijk geeft nog nekpijn aan na een jaar follow-up. Wanneer mensen met nekpijn niet herstellen in de eerste paar weken, leidt de prognose bij een substantiële proportie hiervan tot persisterende of intermitterende pijn en beperkingen in het dagelijks functioneren.
Identificatie van patiënten die op korte termijn zeer waarschijnlijk zullen herstellen kan het risico van overbehandeling en zorg gerelateerde kosten reduceren. Vroegtijdige identificatie van nekpijn patiënten met een te verwachten slechte uitkomst stelt clinici bovendien in staat om tijdig efficiënte behandelingen in te zetten, die de last voor patiënten en de zorg gerelateerde kosten kunnen doen verminderen.
Patiënten met nekpijn kunnen zich zorgen maken om hun toekomst en willen graag hun prognose weten bij het bezoek aan de eerstelijns clinicus. De prognose van het herstel van niet-specifieke nekpijn is een uitdagende taak voor de clinicus. Prognostische factoren en prognostische voorspelmodellen kunnen de clinicus van aanvullende informatie voorzien om de inschatting te verbeteren van de individuele prognose van patiënten. Prognostische factoren geven informatie over de prognose van mensen met nekpijn in het algemeen, terwijl een accuraat prognostisch voorspelmodel de voorkeur heeft omdat deze het voordeel biedt van een individuele prognose voor een specifieke patiënt. Prognostische voorspelmodellen worden voor het geven van een individuele prognose door clinici in de diverse domeinen en settingen van de gezondheidszorg gebruikt en zouden bruikbaar kunnen zijn voor de prognose van niet-specifieke nekpijn in de eerstelijnsgezondheidszorg. Studies naar prognostische voorspelmodellen beslaan drie opeenvolgende stadia: modelontwikkeling (derivatie), bij voorkeur met interne validering; validering in nieuwe settingen (externe validering); en het beoordelen van de klinische impact van een model. De verandering naar persoonlijke zorg (perzonalized medicine) heeft geleid tot de publicatie van een aanzienlijk aantal prognostische voorspelmodellen.

Het algemene doel van dit proefschrift is het verbeteren van het voorspellen van niet-specifieke nekpijn bij individuele patiënten in de eerstelijnsgezondheidszorg met behulp van prognostische voorspelmodellen.

Hoofdstuk 1 is een introductie op het onderzoek dat in dit proefschrift is uitgevoerd en de twee onderzoeksvragen die geëvalueerd werden om dit doel te bereiken. De twee onderzoeksvragen waren:

1. "Zijn er valide prognostische voorspelmodellen beschikbaar om accurate voorspellingen te doen van het herstel van patiënten met niet-specifieke nekpijn?" (**Hoofdstuk 2 en Hoofdstuk 3**). Gebaseerd op het onderzoek dat is verzameld voor deze hoofdstukken, worden veel voorkomende methodologische tekortkomingen in onderzoek naar prognostische voorspelmodellen en aanvullende problemen op het terrein van wervelkolompijn beschreven in **Hoofdstuk 4**.
2. "Kunnen nieuw te ontwikkelen prognostische voorspelmodellen accurate voorspellingen over herstel van nekpijn in de eerstelijnsgezondheidszorg doen, bij patiënten met niet-specifiek nekpijn?" (**Hoofdstuk 5 en Hoofdstuk 6**).

In **Hoofdstuk 2** werd een systematische review uitgevoerd om een overzicht te krijgen van bestaande multivariabele prognostische voorspelmodellen voor het herstel van mensen met niet-specifieke nekpijn die gebruikt zouden kunnen worden in de eerstelijnsgezondheidszorg.
Studies die pijnvermindering, vermindering in beperkingen in het dagelijks functioneren of ervaren herstel als uitkomst hadden, werden geïncludeerd. Ieder follow-up moment werd meegenomen. Er werden drieënvijftig publicaties geïncludeerd, waarvan 46 derivatie studies, vier validering studies en drie combinatie studies. De kwaliteit van de geselecteerde studies werd geëvalueerd met de 'Prediction model study Risk of Bias Assessment Tool' (PROBAST). De PROBAST is een nieuwe tool die ontwikkeld is om het risico op bias, de toepasbaarheid en bruikbaarheid te beoordelen, van primaire studies naar multivariabele voorspelmodellen in een systematische review.
Er werden 99 voorspelmodellen geïdentificeerd die diverse uitkomstmaten van herstel bij mensen met nekpijn voorspelden. Deze voorspelmodellen werden verkregen in 49 studies die allen een hoog risico op bias hadden. Dit was vooral gerelateerd aan de flow van de proefpersonen, de analyses, en verkeerde overwegingen met betrekking tot de steekproefgrootte. De rapportage en de methodologische standaarden waren vaak suboptimaal ten aanzien van performance maten, de omgang met missende data en overfitting.
Zeven voorspelmodellen werden onderworpen aan externe validering, vier daarvan in studies met een hoog risico op bias. Drie voorspelmodellen werden geëvalueerd op 6 of 12 maanden follow-up in externe validering studies met een laag risico op bias: Twee 'Whiplash-Associated Disorder' (WAD) voorspelmodellen die de beperkingen in het dagelijks functioneren voorspellen (met behulp van de 'Neck Disability Index' (NDI) op respectievelijk 6 en 12 maanden follow-up) en een voorspelmodel voor niet-specifieke nekpijn die het ervaren herstel voorspelt op 6 maanden follow-up. Deze drie modellen lijken veelbelovend voor klinisch gebruik. Echter, de WAD voorspelmodellen werden geëvalueerd in cohorten die niet uitsluitend proefpersonen uit de eerstelijnsgezondheidszorg bevatten.
Het voorspelmodel voor niet-specifieke nekpijn had een beperkt discriminerend vermogen ('Area Under de Curve' (AUC) van 0.65 (95% CI 0.59 tot 0.71)). We concludeerden dat deze drie voorspelmodellen verdere validering in een eerstelijnsgezondheidszorg setting nodig hadden, voordat hun klinisch gebruik kon worden aanbevolen.

In **Hoofdstuk 3** wordt de voorspellende prestatie in termen van discriminerend vermogen en kalibratie van deze drie veelbelovende voorspelmodellen geëvalueerd in een externe validering studie. De bevindingen van deze studie werden gerapporteerd volgens de TRIPOD-aanbevelingen. Data van de ANIMO studie werd gebruikt. ANIMO is een prospectieve cohortstudie die uitgevoerd werd van oktober 2007 tot maart 2008 en had tot doel om de standaardbehandeling met manuele therapie voor patiënten met nekpijn te beschrijven in Nederland en uitkomsten en bijwerkingen van behandelingen te verkennen. Het betrof een groot cohortonderzoek (n=1193) waarbij mensen met nekpijn werden gerekruteerd en behandeld door direct toegankelijke manueel therapeuten in Nederland. De standaardbehandeling met manuele therapie kon bestaan uit specifieke gewrichtsmobilisaties en manipulaties, myofasciale technieken, het geven van advies of het uitvoeren van specifieke oefeningen. De prestaties van de voorspelmodellen in termen van discriminerend vermogen en kalibratie bleek slecht voor een van de WAD voorspelmodellen en het niet-specifieke nekpijn voorspelmodel. De AUC-waarden kwamen substantieel onder de 0,7 en de kalibratie slopes weken ruim af van 1,0. Het andere WAD voorspelmodel kon niet worden geëvalueerd daar meerdere variabelen die in het model zijn opgenomen, of hun proxies, niet aanwezig waren in de ANIMO dataset. We concludeerden dat de externe validering van de voorspelmodellen die aanvankelijk veelbelovend leken, niet succesvol was en dat hun klinisch gebruik niet kan worden aanbevolen.

In **Hoofdstuk 4** worden methodologische tekortkomingen in onderzoek naar prognostische voorspelmodellen beschreven. Tijdens ons onderzoek naar prognostische voorspelmodellen, bleken de veel voorkomende methodologische tekortkomingen in het domein van de gezondheidszorg ook zeer veel voor te komen op het terrein van wervelkolompijn. Gebruikelijke methodologische tekortkomingen die ook worden geïdentificeerd op het terrein van wervelkolompijn zijn te veel kandidaat predictor categorieën in relatie tot het aantal events, selecteren van voorspellers op basis van statistische significantie, categoriseren van continue voorspellers, gebrek aan rapporteren van de belangrijkste prestatie maten en slechte algehele rapportage.
Deze veel voorkomende tekortkomingen leiden vaak tot overfitte, overoptimistische of instabiele voorspelmodellen. Dit resulteert in voorspelmodellen die slecht generaliseerbaar zijn naar andere klinische settingen en patiënten. Deze tekortkomingen kunnen voor een groot deel geadresseerd worden door de huidig beschikbare methodologische standaarden te volgen voor het ontwerpen, uitvoeren en rapporteren van voorspelmodellen in de gezondheidszorg.
Daarnaast kwamen we aanvullende methodologische uitdagingen tegen die specifiek waren voor het terrein van wervelkolompijn. Op basis van de evidentie die we verkregen tijdens ons onderzoek, werden vijf aanvullende methodologische problemen geëvalueerd en besproken. Het eerste probleem betrof de keuze van de proefpersonen. Verschillen tussen studies in de selectie van patiënten kan resulteren in case mix verschillen tussen voorspelmodellen die dan moeilijk te vergelijken en interpreteren zijn. Een heldere beschrijving van criteria voor het rekruteren en selecteren, met afbakening van subgroepen die naar verwachting verschillen in prognose (v.b. graad 3 WAD), zou dit probleem tegen kunnen gaan. Het tweede probleem betrof de opzet van de studie. Het is gangbaar dat voor studies naar prognostische voorspelmodellen data wordt gebruikt van andere studies die

focusten op een ander doel. Dit resulteert vaak in het ontbreken van in potentie relevante predictor variabelen en variabelen die niet adequaat worden geoperationaliseerd. Dit probleem kan worden tegengegaan door het gebruik van grote data sets die specifiek worden verzameld voor studies naar prognostische voorspelmodellen en een grote verscheidenheid kennen aan potentieel relevante biologische, fysieke en psychosociale kandidaat voorspellers. Het derde probleem betrof beperkingen in het meten van uitkomstmaten en voorspellers. Deze worden vaak gemeten met door de patiënt gerapporteerde uitkomstmaten (PROMs). Er blijkt een grote verscheidenheid aan diverse PROMs met verschillende afkapwaarden te worden gebruikt. Het ontwikkelen van een op consensus gebaseerde kern set van uitkomstmaten kan een oplossing zijn hoe om te gaan met deze heterogeniteit. Beperkingen in het meten zoals insufficiënte constructvaliditeit, inhoudsvaliditeit en structurele validiteit, kunnen de prestaties van PROMs in een prognostisch voorspelmodel beïnvloeden. Verder beïnvloeden meetfouten van PROMs de prestatie van voorspelmodellen; random fouten verminderen de kalibratie en het discriminerend vermogen, terwijl systematische fouten kalibratie beïnvloedt en niet het discriminerend vermogen. Het vierde probleem betrof de complexiteit van het voorspellen van herstel. Niet-specifieke wervelkolomproblemen kunnen typisch beschouwd worden als complexe gezondheidsproblemen waarbij vele interacterende factoren bijdragen aan de prognose van pijn en beperkingen in het dagelijks functioneren. De huidige aanpak in het maken van voorspelmodellen pakt de nog altijd vele onbekende variabelen en hun interacties niet op, die in de tijd ook dynamisch kunnen veranderen. Het opnemen van interactie tussen variabelen en variabelen gebaseerd op trajecten in het verloop van de tijd, hebben potentie om de prestatie van voorspelmodellen te verbeteren. Kunstmatige intelligentie en machine learning technieken zouden in staat moeten zijn om nieuwe voorspellers te ontdekken en te modelleren in grote datasets, waarbij data van diverse bronnen gekoppeld kunnen worden. Het vijfde probleem betrof de verwarring van prognose en behandelrespons. Prognostische factoren zijn niet altijd direct ook een voorspeller van het behandeleffect. Voorspellers van het behandeleffect worden geëvalueerd door de interactie van die predictor met de behandeling, als additioneel effect op de uitkomstmaat, te onderzoeken. Eenarmige cohortstudies leveren verkennende informatie op. Er zijn echter tweearmige trials nodig voor hun sterkere ontwikkeling en validering.

De bevinding van slecht presterende voorspelmodellen in de externe validering studie (**Hoofdstuk 3**), bracht ons op de tweede onderzoeksvraag of een nieuw te ontwikkelen prognostisch voorspelmodel accurate voorspellingen over herstel van nekpijn in de eerstelijnsgezondheidszorg kan doen, bij patiënten met niet-specifieke nekpijn. We besloten een nieuw prognostisch voorspelmodel voor herstel te ontwikkelen en valideren, waarbij we nadrukkelijk de veel voorkomende methodologische tekortkomingen voorkwamen.
Het doel van **Hoofdstuk 5** was het ontwikkelen en intern valideren van modellen voor het voorspellen van herstel bij patiënten met niet-specifieke nekpijn. Hiervoor werd de data van het eerder beschreven grote ANIMO cohortonderzoek gebruikt (n=1193), waarbij mensen met nekpijn werden gerekruteerd en behandeld door manueel therapeuten in Nederland. De uitkomstmaten die gebruikt werden om het herstel te definiëren waren de pijnintensiteit, nekpijn gerelateerde beperkingen in het dagelijks functioneren, en de globaal ervaren verbetering, direct na de behandelingen en op 1 jaar follow-up.
Veertien tot achttien categorieën van kandidaat voorspellers werden meegenomen in de

multivariabele analyses voor de zes voorspelmodellen. Een discriminerend vermogen van AUC ≥ 0,70 werd als acceptabel beschouwd. Het voorspelmodel dat herstel van beperkingen in het dagelijks functioneren (verder als disability model) direct na de behandelingen voorspelt, vertoonde de beste overall prestatie R^2=0,24 (IQR, 0,22–0,26), discriminerend vermogen AUC=0,75 (95% CI: 0,63–0,84), en kalibratie (slope 0,92; interkwartiel range (IQR): 0,91–0,93). Na interne validering, vertoonde het voorspelmodel een discriminerend vermogen van AUC=0.74 (IQR: 0,72-0,75). Hiermee had dit voorspelmodel de beste potentie voor klinisch gebruik. De andere twee modellen voor het voorspellen van herstel van pijn en de globaal ervaren verbetering direct na de behandelingen, vertoonden initieel ook acceptabele prestaties, maar behielden deze niet na interne validatie. Geen van de ontwikkelde voorspelmodellen die het herstel op 1 jaar voorspellen, bereikte een acceptabele prestatie.

In **Hoofdstuk 6** voerden we een breed extern validering onderzoek uit in de data van het PRONEPA-cohortonderzoek van de prognostische voorspelmodellen die waren ontwikkeld in **Hoofdstuk 5**. De voorspelmodellen werden geëvalueerd direct na de behandelingen en op 6- en 12 weken follow-up. PRONEPA is een prospectief cohortonderzoek die in Nederland uitgevoerd werd van November 2020 tot April 2021, die primair tot doel had prognostische factoren te evalueren die het ontwikkelen van chronische nekpijn voorspellen bij mensen met (sub)acute nekpijn (n=586). De proefpersonen werden gerekruteerd uit personen die zich aangemeld hadden bij eerstelijns fysiotherapiepraktijken en werden behandeld door fysiotherapeuten die een Master of Science opleiding in manuele therapie volgden. De behandelingen bestonden uit standaardbehandeling fysiotherapie, gebaseerd op de richtlijn. Een discriminerend vermogen van AUC ≥ 0,70 werd als acceptabel beschouwd. De externe validering van het disability voorspelmodel dat herstel na 6 weken voorspelt, vertoonde een discriminerend vermogen van AUC=0,73 (95% CI: 0,69-0,77) en een behoorlijk goede kalibratie na re-kalibratie van de intercept. Externe validatie van dit voorspelmodel op 12 weken en direct na de behandelingen vertoonde een bijna acceptabel discriminerend vermogen van respectievelijk AUC 0,69 (95% CI: 0,64-0,73) en 0,68 (95% CI: 0,63-0,72) en waren goed gekalibreerd.
Daarnaast waren we geïnteresseerd of de voorspelmodellen die we in **Hoofdstuk 5** verkregen hadden, konden worden geüpdatet met beschikbare variabelen in de data van het PRONEPA-cohort, met name cervicale mobiliteit, uithoudingsvermogen van de anterieure nekspieren en pijn catastroferen. Cervicale mobiliteit bleek voor alle follow-up momenten significant bij te dragen aan het disability voorspelmodel en pijn catastroferen ook aan het 6-weken pijn voorspelmodel.
Deze toegevoegde predictoren verbeterden de prestaties van de voorspelmodellen echter minimaal. We stellen voor dat fysiotherapeuten het disability voorspelmodel, zonder de toegevoegde predictoren, tijdens de intake gebruiken voor de prognose van mensen met nekpijn, om hen te assisteren in klinische beslissingen over herstel van beperkingen in het dagelijks functioneren na 6 weken. Vervolgonderzoek is nodig om de klinische impact van het disability voorspelmodel te beoordelen.

Tot besluit reflecteren we in **Hoofdstuk 7** op de belangrijkste bevindingen van dit proefschrift, op het stellen van prognoses in de klinische praktijk, en staan we stil bij de implicaties voor de klinische praktijk en wetenschappelijk onderzoek.

Dankwoord

Toen ik direct na mijn universitaire studie een promotie traject opperde, reageerde mijn gezin resoluut met een “nee, dan zien we je helemaal nooit meer!” en was het mij volledig duidelijk dat ik hier toen niet aan moest beginnen. Enkele jaren later was hun reactie een heel andere en met een “natuurlijk doen pap” en een “ik zie je ogen glimmen bij het idee” besloot ik deze uitdagende stap te wagen. Vanaf dat moment vulde ik het werken in onderwijs en in mijn klinische praktijk aan met het uitvoeren van wetenschappelijk onderzoek, wat voor mij een bijzonder inspirerende en lerende kruisbestuiving werd. Ik wil dan ook beginnen met een algemeen dankwoord aan iedereen in het onderwijs, praktijk en onderzoek die mij geholpen en geïnspireerd heeft. Zonder iemand tekort te willen doen, bedank ik de volgende mensen graag ook nog persoonlijk:

Prof. Dr. B. W. Koes, promotor. Beste Bart, bedankt dat ik onder jou mocht promoveren bij de afdeling huisartsgeneeskunde van het Erasmus MC. De gastvrijheidsovereenkomst bleek daadwerkelijk te staan voor een gastvrije afdeling. Met een laagdrempelige en positieve begeleiding bracht je mijn soms wat uitweidende manuscripten altijd weer met enkele rake en nuchtere opmerkingen terug tot de essentie.

Dr. E. van Trijffel, inspirator en aanvankelijke copromotor die mij vanaf de eerste stap in het promotietraject inspireerde en vertrouwen gaf. Beste Emiel, de kwaliteit, toonzetting en diepgang van jouw begeleiding is indrukwekkend. Iedere keer bracht je me met heldere en scherpzinnige feedback op weer een ander level. Ook nadat het carrière pad je buiten de SOMT bracht bleef je onvermoeid en snel je wijze feedback op mijn producten geven. Bedankt dat je altijd maar weer op deze manier voor mij klaar stond.

Prof. Dr. A.P. Verhagen, aanvankelijke copromotor die gedurende dit promotie traject de grote sprong naar Australië maakte. Beste Arianne, je hebt mij flink op weg geholpen in de wereld van prognostische modellen. Hoewel geen Rotterdammer, gaf je lekker directe feedback waar ik uitstekend mee uit de voeten kon. Bedankt dat je mij met je geweldige onderzoekservaring hebt willen begeleiden.

Dr. A. Chiarotto, copromotor. Beste Alessandro, er werd me weleens gevraagd of ik de wisselingen in het begeleidingsteam niet lastig vond, maar ook met jouw begeleiding trof ik het weer enorm. Bedankt voor je constructieve en zorgvuldige inbreng en bedankt dat je na Emiel en Arianne het stokje van de copromotie zo probleemloos overnam.

Dr. M. W. Heymans. Beste Martijn, je was al vroeg in het promotie traject betrokken vanwege je statistische deskundigheid en je werd snel een vast onderdeel van het team. De wereld van prognostische modellen raakt aan verassend veel aspecten van statistiek. Bedankt voor je adviezen en wijze raad bij al deze statistische aangelegenheden.

Mijn medeauteurs, Drs. Paul M. Nelissen, Dr. Martijn S. Stenneberg en Drs. Ronald Kan, wil ik graag bedanken voor hun bijdrage aan het eerste en laatste artikel. Beste Paul, bedankt voor je waardevolle samenwerking bij de systematische review.

Fijn dat we na weer een dag hard werken, samen toch nog kans zagen even te gaan hardlopen. Beste Martijn en Ronald, met jullie hulp konden we ons model valideren in de data van de PRONEPA-studie, bedankt voor jullie belangrijke inbreng en bijdrage aan de validatie studie. Als dagelijkse collega’s gaan we vast nog meer mooie onderzoeken doen.

Graag wil ik de leden van de leescommissie, Prof. dr. ir. A. Burdorf, Prof. dr. B. Cagnie en Prof. dr. C. Lucas, hartelijk danken voor hun bereidheid het manuscript voor dit proefschrift te beoordelen en voor hun oppositie. Ook bedank ik graag de overige leden van de promotie-commissie, Prof. dr. G.M. Ribbers en Dr. Scholten-Peeters voor hun aanwezigheid en hun oppositie.

Wat fijn dat ik deze twee paranimfen tijdens mijn verdediging naast me weet.
Drs. R. Wingbermühle, lieve Robin, wat heerlijk dat jij als oudste de drie dochters (en al ietsje meer) wil vertegenwoordigen. E. L. Bol, beste Erwin, wat fantastisch dat jij als praktijkmaat en vriend nu naast me staat. Sorry dat je me de afgelopen jaren zoveel tijd hebt moeten missen.

Drs. W. Smeets, beste Willy, hartelijk bedankt voor de gelegenheid die je mij bood om te promoveren. Dankzij jouw visie draagt de SOMT en haar team een wezenlijk steentje bij aan wetenschappelijk onderzoek en aan academisch onderwijs op het terrein van de fysiotherapie.

Natuurlijk wil ik ook Aad van de El, alle collega’s en oud-collega’s van de SOMT bedanken. Na al die vele jaren sinds 1991 zou dit een veel te lange lijst met namen worden om op te noemen, met het aanzienlijke risico dat ik iemand zou vergeten te noemen. Ik ben enorm dankbaar dat ik met jullie mag en mocht samenwerken en dat ik zoveel vreselijk veel van jullie heb kunnen leren. Wat bijzonder dat we als team altijd zoveel voor elkaar over hadden en hebben.

Beste praktijkcollega’s, onderwijs en onderzoek hebben mij door de jaren heen steeds meer weggevoerd uit de praktijk. Ik realiseer me terdege dat niet iedereen mijn afwezigheid altijd even leuk vindt. Wat heerlijk dat ik toch nog altijd het onderzoek bij ons in de praktijk mag blijven spiegelen.

Beste Patty, bedankt voor al je hulp bij de opmaak en je geduld bij het tot stand komen van het uiteindelijke boekje.

Lieve pa en ma, Eef en Ton, familie, vrienden, functionele beesten en amici, bedankt voor al jullie interesse in mijn werk en de vorderingen. Wat geweldig dat jullie mij ook tijdens al onze gezamenlijke vakanties (en autoritten daarnaartoe), regelmatig de gelegenheid boden achter mijn laptop weg te duiken om weer eens door te kunnen werken aan mijn onderzoeken en manuscripten. Pa, wat jammer dat je mijn uiteindelijke promotie niet meer mee kon maken. Ik draag dit proefschrift dan ook op aan jou.

Robin & Timo, Luca en Noa, geweldig dat jullie me ook na die eerste stap nog zijn blijven steunen.

Marjan, mijn meisje, mijn lief. Wat een heerlijk besef en gevoel dat we elkaar altijd maar weer de ruimte gunnen om nieuwe uitdagingen aan te gaan en dat we elkaar daarbij steeds weer willen blijven vinden.

Curriculum vitae

Roel Wingbermühle was born on September 3, 1963, in Amersfoort, the Netherlands. He is son of Jacques Wingbermuhle and Ria van de Kamp. He grew up in Krimpen aan den IJssel with his sisters Iris and Manon. He is married to Marjan Olyslager and lives in Rotterdam. They have three daughters Robin, Luca and Noa and a son-in-law Timo.

After graduating from Emmaus College, Roel studied Physiotherapy in Rotterdam. He graduated in 1984 and fulfilled his military service as a physiotherapist in Roosendaal. Thereafter, he started working in a primary care physiotherapy practice in Krimpen aan den IJssel, nowadays called Fysiotherapie MSC Wingbermühle | Bol.
After his study of manual therapy at SOMT Eindhoven, he started teaching manual therapy at this same institution in 1991. This led him to teach manual therapy in Germany, Switzerland, and Belgium as well.

In 2010 he obtained his Master of Science in manual therapy (cum laude) at the Free University Brussels, Belgium. In 2015 he started as a PhD student at the Erasmus MC, department of general practice. His promotor was prof. Dr. Bart Koes and co promotors Dr. Arianne Verhagen and Dr. Emiel van Trijffel, followed by Dr. Alessandro Chiarotto. From this moment, he combined scientific research with working at SOMT and his primary care physiotherapy practice. At present, he is head of the division of bachelor education at SOMT.

List of publications

Wingbermühle R.W., van Trijffel E., Nelissen P.M., Koes B.W., Verhagen A.P.
Few promising multivariable prognostic models exist for recovery of people with non-specific neck pain in musculoskeletal primary care: a systematic review.
Journal of Physiotherapy (2018) 64(1):16-23

Wingbermühle R.W., Heymans M., van Trijffel E., Chiarotto A., Koes B.W., Verhagen A.P.
External validation of prognostic models for recovery in patients with neck pain.
Brazilian Journal of Physical Therapy (2021) 25(6):775-784

Wingbermühle R.W., Chiarotto A., Koes B.W., Heymans M., van Trijffel E.
Challenges and solutions in prognostic prediction models in spinal disorders.
Journal of Clinical Epidemiology (2021) 132:125-130

Wingbermühle R.W., Chiarotto A., van Trijffel E., Koes B.W., Verhagen A.P., Heymans M.
Development and internal validation of prognostic models for recovery in patients with non-specific neck pain presenting in primary care. Physiotherapy (2021) 113:61-72

Wingbermühle R.W., Chiarotto A., van Trijffel E., Stenneberg M.S., Kan R., Koes B.W., Heymans M. External validation and updating of prognostic models for predicting recovery of disability in people with (sub)acute neck pain was successful: broad external validation in a new prospective cohort. Submitted.

Acknowledgements
Dr. E. van Trijffel, rijff037@planet.nl
Dr. A. Chiarotto (co-promotor), a.chiarotto@erasmusmc.nl
Prof. Dr. A. P. Verhagen, arianne.verhagen@uts.edu.au
Dr. M.W.H. Heijmans, mw.heymans@amsterdamumc.nl
Prof. Dr. B. W. Koes (promotor), b.koes@erasmusmc.nl

PhD Portfolio

Erasmus MC Department: General Practice
PhD Period: September 2015 – July 2022
Promotor: Prof. dr. B.W. Koes
Co-promotor: dr. A. Chiarotto

	Year	Workload ECTS
Courses/training		
Masterclass anderhalvelijns fysiotherapie	2015	4.0
Basiskwalificatie onderwijs (BKO), Maastricht University	2015	1.0
Clinical prediction models, Maastricht University	2016	1.0
Clinical prediction models, EpidM, VU Amsterdam	2018	1.0
Basiskwalificatie Examinering (BKE), SOMT Amersfoort	2018	1.0
Kwalitatieve onderzoeksmethoden, Universiteit Antwerpen	2019	1.0
Scientific Integrity, Erasmus MC, Rotterdam	2022	0.3
Oral presentations		
Back and neck pain forum	2021	0.5
Poster presentations		
DvdF poster presentation	2017	0.25
Memtab congress	2018	0.5
Teaching & Career		
Teaching activities and BSc program development	2015-2020	166
Head of WO bachelor education	2020-2021	53
Total		229.5