Journal of Physiotherapy ■ (2019) ■–■

AUSTRALIAN PHYSIOTHERAPY ASSOCIATION

Journal of PHYSIOTHERAPY

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

Appraisal

Research Note: Prognostic model research: overfitting, validation and application

Introduction

In physiotherapy, many prognostic models have been developed to predict future outcomes after musculoskeletal conditions, including neck pain.[1] Prognostic models combine several characteristics to predict the risk of an outcome for individual patients and may enable personalised prevention and care. In practice, they can be used to inform patients and relatives on prognosis, and to support clinical decision-making. Moreover, models may be useful to stratify patients for clinical trials. Prediction models are increasingly being published, including 99 prognostic models for neck pain alone, predicting recovery (pain reduction, reduced disability, and perceived recovery).[2] Although guidelines for developing and reporting prognostic models have been proposed,[3,4] a recently proposed assessment tool found that many prognostic models in physiotherapy are prone to risks of bias.[2,5]

Various limitations have been noted regarding design and analyses, which make models at risk of overfitting.[2] Overfitting relates to the notion of asking too much from the available data, which results in overly optimistic estimates of model predictive performance; results that cannot be validated in underlying or related populations.[6] Consequently, the model may predict poorly, with serious limitations when the model is applied in clinical practice: it does not separate low-risk from high-risk patients (poor discrimination), and may give unreliable or even misleading risk estimates (poor calibration).

We aim to describe a number of challenges related to the design and analysis in different stages of prognostic model research, and opportunities to reduce overfitting (summarised in Table 1). We emphasise validation before the application of prediction models is considered in clinical practice. For illustration, we consider the Örebro Musculoskeletal Pain Screening Questionnaire (OMPQ) (Table 2).[7] The model has extensively been validated, and its use is recommended by clinical guidelines.[8] We also consider the Schellingerhout non-specific neck pain model predicting recovery after six months (Table 2),[9] which was indicated as one of the few externally validated models with a low risk of bias.[2]

Model development

The development of a prognostic model involves a number of steps. These include handling of missing data, selection and coding of predictor variables, choosing between alternative statistical models, and estimating model parameters.[10] Prognostic models are usually developed with multivariable regression techniques on data from (prospective) cohort studies, while machine learning techniques are gaining increased attention.

Missing data is common in prognostic research. A complete case analysis is often conducted (ie, the exclusion of participants that have missing data on one or multiple predictor variables, resulting in smaller sample size). As a consequence, the number of events per variable may drop below the number deemed necessary for reliable modelling (Table 1), increasing the risk of overfitting. Better approaches are imputation methods,[10] where missing values may be substituted with the mean or the mode with single imputation, and *m* completed data sets are created with multiple imputation procedures. Multiple imputation is recommended, because single imputation ignores potential correlation of predictors and leads to an underestimation of variability of predictor values among subjects.[11] This may lead to an overestimation of the precision of regression coefficients. Imputation methods are widely available through modern statistical software.

It is difficult to select the most promising predictors. Selection of candidate predictors based on literature and expert knowledge is often preferred over selection based on a relatively limited dataset.[10] Also, some related predictors can sometimes be combined in simple scores. For example, comorbid conditions are often combined in a comorbidity score,[12] and frailty in the elderly can be scored according to various characteristics.[13] After selection of candidate predictors, the set of predictors may be reduced; this can be done using univariate analysis and/or stepwise methods. However, both approaches do not truly reduce the problem of statistical overfitting, since the model specification is driven by findings in the data. Univariate analysis is common as a first step to select the most potent risk factors, which are then used in multivariable analysis. This approach was followed in the development of the OMPQ (Table 2). A common alternative is to use backward stepwise selection from a model that includes all candidate predictors, as was done by Schellingerhout to develop a model to predict non-specific neck pain (Table 2). Stepwise selection procedures are known to result in biased regression coefficient estimates (*testimation bias*).[6] A modern approach to reduce such *testimation bias* and overfitting is by shrinkage of regression coefficients towards zero.[10] A key example of this approach is the Least Absolute Shrinkage and Selection Operator, which penalises for the absolute values of the regression coefficients. It shrinks some coefficients to zero, which means that predictors are dropped from the model.

Validation: apparent, internal and external performance

The aim of prognostic models is to provide accurate risk predictions for new patients. Therefore, validation of prognostic models is crucial. Three types of validation can be distinguished: apparent, internal and external validation.

Apparent validation entails the assessment of model performance directly in the derivation cohort. Because the regression coefficients are optimised for the derivation cohort, this provides optimistic estimates of the model's performance (overfitting). To correct for overfitting, several internal validation procedures are available. Bootstrap resampling and cross-validation provide stable estimates with low bias and are therefore recommended.[10]

Before a prognostic model can be applied in practice it is crucial to explore how the model performs outside the setting in which it was developed, preferably across a range of settings. External

https://doi.org/10.1016/j.jphys.2019.08.009

Table 1
Overview of challenges and opportunities categorised by the stage of prognostic model research in which they occur, and illustrated with two prediction models.[7,9]

Stage of prognostic model research	Challenges	Opportunities	Örebro Musculoskeletal Pain Screening Questionnaire	Schellingerhout non-specific neck pain model
Design	Insufficient sample size	Collaborative efforts to reach > 10 EPV, cross-validate across setting	No information on EPV	Restricted to 17 predictors based on EPV (10)
Development	Inappropriate handling of missing data; complete case analysis	Multiple imputation methods	Complete case analysis	Multiple imputation with 5 repetitions
Development	Selection of predictors based on univariate analysis or stepwise selection procedures	Shrinkage and penalisation in multivariable analysis	Univariate analysis	Backward stepwise selection
Internal validation	Apparent validation or inefficient internal validation procedures	Bootstrap resampling or cross-validation	Apparent validation	Apparent validation
External validation	Full model equation is not presented	Present full model equation	Yes	Yes
External validation	No external validation	Validation of models in cohort other than development cohort through collaborative research	Externally validated; AUC, but no calibration plot	Externally validated; AUC and calibration plot

EPV = events per variable; AUC = area under the receiver operating characteristic curve.

validity relates to the generalisability of the prognostic model to another population.[10] A cross-validation across different non-random parts of the development data gives an indication of external validity.[14] Heterogeneity in predictor effects across settings indicates that the model should be calibrated to each specific setting, to achieve robust model performance across settings. To enable external validation of the model the full model equation should be presented in the paper (Table 1). The OMPQ has been extensively validated in international cohorts,[15] while such external validation is rare for other prognostic models for musculoskeletal conditions.[2,16]

Performance measures

Model performance at internal and external validation is commonly expressed with discrimination and calibration. Discrimination indicates the ability of the model to differentiate between high-risk and low-risk patients. It can be measured by the concordance statistic (C-statistic or area under the receiver operating characteristic curve: AUC). The AUC ranges between 0.50 (no discrimination) and 1.00 (perfect discrimination). For instance, the OMPQ was validated in an observational study of patients with acute back pain in Australia.[17] At external validation of the OMPQ the AUC was 0.80 (95% CI 0.66 to 0.93) for absenteeism at 6 months (Table 2).[17] The discriminative ability of the Schellingerhout non-specific neck pain model was lower: AUC 0.66 (95% CI 0.61 to 0.71) at development, and validation cohort AUC 0.65 (95% CI 0.59 to 0.71).[9]

Calibration refers to the agreement between predicted and observed probabilities. This agreement can be illustrated with a calibration graph. Ideally, the plot shows a 45-deg line with calibration slope 1 and intercept 0. Calibration is more informative at external than internal validation because a model is expected to provide correct predictions for the derivation cohort it is fitted on. At external validation, the Schellingerhout non-specific neck pain score chart showed reasonable calibration (Figure 1); it slightly

Table 2
Overview of prognostic model characteristics of the Örebro Musculoskeletal Pain Screening Questionnaire and the Schellingerhout non-specific neck pain model.

	Örebro Musculoskeletal Pain Screening Questionnaire	Schellingerhout non-specific neck pain model
Development		
Patient population of development cohort	n = 137; adult patients; acute/subacute back pain; Sweden[7]	n = 468; adult patients (18 to 70 yrs); non-specific neck pain; primary care; The Netherlands[9]
Outcome	Accumulated sick leave; 6 months follow-up	Global perceived recovery; dichotomised into 'recovered or much improved' versus 'persistent complaints'; 6 months follow-up
Predictors	21 predictors: physical functioning, fear-avoidance beliefs, the experience of pain, work, and reactions to the pain	9 predictors: age, pain intensity, previous neck complaints, radiation of pain, accompanying low back pain, accompanying headache, employment status, health status, and cause of complaints
External validation		
External validation	n = 106; adult patients; acute/subacute low back pain; workers' compensation and medical practitioner referral; observational study; Australia[17]	n = 346; adult patients (18 to 70 yrs); non-specific neck pain; primary care; randomised controlled trial; PANTHER trail; United Kingdom[9]
Model performance	AUC 0.80 (CI 95% 0.66 to 0.93); no calibration plot	AUC 0.65 (CI 95% 0.59 to 0.71); calibration plot
Application		
Practical application	Recommended in clinical guidelines as screening instrument,[8] and used to select trial participants[19]	Score chart[9]

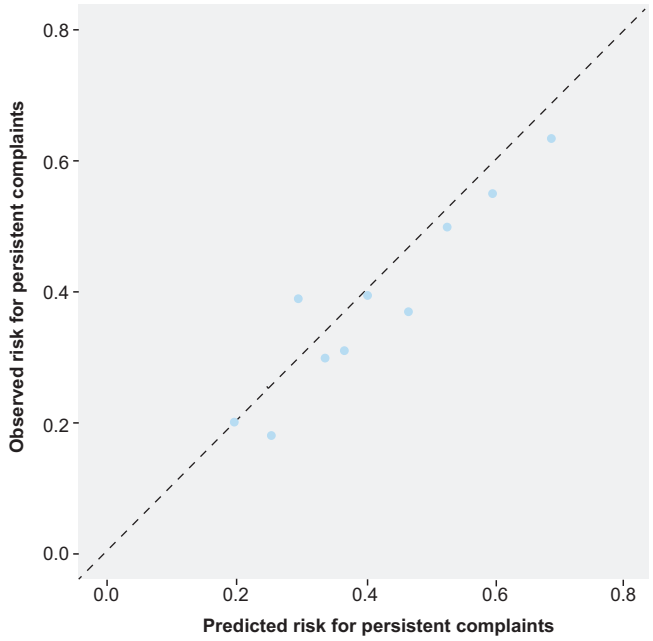

Figure 1. Calibration of the Schellingerhout non-specific neck pain score chart in external validation cohort.
• = deciles of risk
—— = Perfect calibration
Adapted from Schellingerhout et al.[9]

overestimated the risk of persistent complaints in adult patients presenting with non-specific neck pain.[9] More severe miscalibration is common for prediction models.[18]

Application of prognostic models in practice

A prognostic model is more likely to be applicable for implementation in practice if the model was developed with high-quality data from an appropriate study design, and with careful statistical analysis.[10] Even better is when the model is externally validated in the setting where it is to be used.[14] For instance, the OMPQ is recommended in clinical guidelines to be applied in screening to predict delayed recovery,[8] and was used to select trial participants,[19] likely motivated by the extensive and positive external validation studies across multiple settings. When a prognostic model is deemed appropriate for implementation, the impact (clinical effectiveness and costs) of the use of the model in clinical practice should be studied.[4] Although recommended, these clinical impact studies are scarce, and some prediction models have been recommended to be used in clinical practice without adequate evaluation of their (cost-)effectiveness.

The presentation of clinical prediction models is important to facilitate implementation of prognostic models in practice. The Schellingerhout model was presented as a score chart that can readily be used by physicians. Although the score chart may be easy to use, predictions of risks are only approximate because continuous predictors are categorised and regression coefficients are rounded. The score chart should ideally be externally validated across various settings before it can be considered for use in broader practice. Other common formats include web-based calculators and apps for mobile devices.[10,20]

Summary

The aim of prognostic models for predicting future outcomes after musculoskeletal conditions is to provide accurate and patient-specific estimates of the risk of relevant clinical outcomes such as delayed recovery. These models may be applied in primary care to identify patients likely to have poor outcomes. Most models in physiotherapy have been judged to be at moderate to high risk of bias.[2] Approaches to reduce overfitting should be better utilised. These include appropriate handling of missing data, careful selection of predictors with domain knowledge, and internal and external validation (Table 1). Assessment of performance across a range of settings may show suboptimal results, specifically with respect to calibration of predictions. Such suboptimal performance may motivate updating of a model before it can be considered for application in a specific setting.[10] Furthermore, clinical impact studies are recommended to assess the (cost-) effectiveness of a prognostic model in clinical practice. The presentation format of a prognostic model is also important, as this can facilitate implementation of prognostic models in clinical practice to improve decision-making and outcome by personalised medicine.

Competing interests: Nil.
Sources of support: Nil.
Acknowledgements: None.
Provenance: Invited. Not peer reviewed.

Correspondence: Isabel RA Retel Helmrich, Public Health, Center for Medical Decision Making, Erasmus MC-University Medical Center Rotterdam, The Netherlands. Email: i.retelhelmrich@erasmusmc.nl

Isabel RA Retel Helmrich[a], David van Klaveren[a,b] and Ewout W Steyerberg[a,c]

[a]*Department of Public Health, Center for Medical Decision Making/ Erasmus MC-University Medical Center Rotterdam, The Netherlands*
[b]*Predictive Analytics and Comparative Effectiveness Center, Institute for Clinical Research and Health Policy Studies/Tufts Medical Center, Boston, USA*
[c]*Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands*

References

1. Kelly J, et al. *Musculoskelet Sci Pract*. 2017:155–164.
2. Wingbermühle RW, et al. *J Physiother*. 2018;1:16–23.
3. Collins GS, et al. *BMC Med*. 2015;1:1.
4. Steyerberg EW, et al. *PLoS Med*. 2013;2:e1001381.
5. Wolff RF, et al. *Ann Intern Med*. 2019;1:51–58.
6. Babyak MA. *Clin J Pain*. 1998;3:209–215.
7. Linton SJ, et al. *Clin J Pain*. 1998;3:209–215.
8. ACC. New Zealand acute low back pain guide. https://www.acc.co.nz/assets/provider/ff758d0d69/acc1038-lower-back-pain-guide.pdf. Accessed 14 June, 2019.
9. Schellingerhout JM, et al. *Spine J*. 2010;17:E827–E835.
10. Steyerberg EW. *Stat Methods Med Res*. 2007;3:277–298.
11. Ambler G, et al. *Stat Methods Med Res*. 2007;3:277–298.
12. Charlson ME, et al. *J Chronic Dis Manag*. 1987;5:373–383.
13. Searle SD, et al. *BMC Geriatr*. 2008;1:24.
14. Steyerberg EW, et al. *J Clin Epidemiol*. 2016:245–247.
15. Hockings RL, et al. *Spine J*. 2008;15:E494–E500.
16. van Oort L, et al. *J Clin Epidemiol*. 2012;12:1257–1266.
17. Linton SJ, et al. *Clin J Pain*. 2003;2:80–86.
18. Riley RD, et al. *BMJ*. 2016:i3140.
19. Schmidt CO, et al. *BMC Musculoskelet Disord*. 2010;1:5.
20. Bonnett LJ, et al. *BMJ*. 2019:l737.