Methodological issues of clinical prediction models for shoulder pain in general practice

Vergouw, D.

2013

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

SUMMARY

In primary care and general practice especially, musculoskeletal complaints are frequently occurring and pose a major burden on health care and society. Managing musculoskeletal disorders in primary care poses difficulties since identifying the exact cause of musculoskeletal complaints in individual patients proves to be problematic. Because of the lack of good diagnostic criteria, research has therefore focussed on exploring the determinants of an unfavourable course of musculoskeletal complaints, rather than trying to find a precise cause. Combining such determinants in a clinical prediction model facilitates a quantitative estimate of the likely future outcome which may subsequently be used by physicians as assistive tools for making treatment decisions or for informing and advising patients. However, finding a parsimonious set of determinants, i.e., predictors, to form a simple yet good model that can consistently be applied in a broad patient population proves to be difficult. Some methodological issues such as missing data, or model stability might hinder the development of a prediction model and therefore remain to be resolved. The principle aim of this thesis was to address several methodological issues of clinical prediction models by applying various modelling techniques in several musculoskeletal datasets in order to contribute to the identification of optimal methods for the development and validation of prediction models.

In *chapter 1* we present the aims and outline of this thesis. Furthermore, we describe the epidemiology of musculoskeletal disorders in general, highlight how prediction models might be beneficial in the management of musculoskeletal complaints in primary care, and emphasize how methodological issues such as missing data and model stability might hamper the derivation of clinically feasible prediction models.

In *chapter 2* we aimed to provide researchers with an empirical illustration of the handling of missing data by complete case analysis (CCA) and multiple imputation (MI). Data came from the Beliefs about Backpain (BeBack) cohort, a study of psychological obstacles to recovery in 1591 primary care back pain patients in the United Kingdom. Patients had to give permission for baseline and follow-up contact separately, resulting in missing data in baseline characteristics (14%) and non-response to follow-up (51%). We observed that patients with missing baseline data and patients with missing follow-up data both differed from patients with complete data regarding the distribution of some predictors and outcome, thus creating a selective but non-informative loss of data. As a result, we showed that models derived by using complete cases only differed from models derived by multiple imputation regarding model composition (i.e., predictors included in the final model), model performance, and overoptimism. Since in the presence of missing data CCA may lead to biased results and MI is known to reduce the risk of biased results, our results illustrate that because of this bias and of its possible clinical consequences, MI is recommended over the use of CCA in the presence of missing data.

Although multiple imputation is recommended in the presence of missing data, it introduces two other troubling factors that might hinder predictive modelling; 1) multiple imputed data sets need to

combined to form a single model, and 2) the extra source of instability introduced by MI (i.e., imputation uncertainty) to the already unstable method of automated variable selection. In *chapter 3* we examined the addition of a bootstrap model selection procedure to a MI approach in order to address the troubling factors introduced by MI. The Dutch Shoulder Study, a cohort of 587 patients consulting with a shoulder problem in general practice in the Netherlands, was used to derive models predicting persistent shoulder disability and models predicting persistent shoulder pain. Using a bootstrap resampling method we first separated the strong from weak predictors and subsequently considered model composition in large numbers of resampled data. By doing so we demonstrated how this bootstrap model selection method provided additional information on model stability in models derived from multiple imputed data sets.

In *chapter 4* we aspired to compare one of the most frequently used alternative modelling techniques; the Classification And Regression Tree (CART) methodology, to the commonly used logistic regression analysis. In order to determine which method was better suited for deriving a prediction model for persistent shoulder pain we applied both methods to the Dutch Shoulder Study (DSS), a cohort of 587 patients consulting with a shoulder problem in general practice in the Netherlands. We compared both CART and logistic regression models at several important steps of model development. Although the total number of included predictors was the same (7 predictors) for both models, model composition differed. Model performance in the original data set showed strong resemblance (equal R2N of 19%, an AUC of 0.72 for logistic regression vs. 0.70 for CART). However, when we applied both models to comparable subjects, the CART model was less (internally) valid than the regression based model. Because of this we conclude that our logistic regression model is

better suited for the prediction of persistent shoulder pain than our CART model.

In *chapter 5* our objective was to establish consensus among clinical experts regarding predictors that are most important for predicting persistent shoulder pain in primary care. Secondary we set out to assess the predictive performance of a model based on this clinical consensus. We formed a multidisciplinary and international (United Kingdom and the Netherlands) panel of 41 experts with thorough knowledge of and expertise in the management of (patients with) shoulder pain in clinical practice. In three consecutive Delphi rounds the expert panel selected the following predictors as key determinants of persisting shoulder pain; symptom duration, pain catastrophizing, symptom history, fear-avoidance beliefs, coexisting neck pain, severity of shoulder disability, multisite pain, age, shoulder pain intensity and illness perceptions. Subsequently we transformed these predictors to two prediction models, one using complete continuous information and one using dichotomized predictors. Using a sample of 587 primary care patients consulting with shoulder pain in primary care we compared the predictive performance of both expert-based models to a previously (*chapter 3*) derived statistical prediction model. We showed that the statistically derived model performed better than the expert-based models (statistical model AUC=0.702 vs. AUC= 0.656 expert-based dichotomous, and AUC=0.679 expert-based continuous predictors). We concluded that although expert-based and statistical models were different, both confirmed the prognostic importance of symptom duration, baseline level of shoulder disability and multisite pain as predictors of persistent shoulder pain three months after initial consultation. Although the statistical model appeared to perform better in comparable subjects, we concluded that external validation in other populations of shoulder pain patients should

confirm whether statistically derived models indeed perform better compared to models based on clinical expertise.

In *chapter 6* we investigated the external validity of the two most promising models, the expert-based continuous model from *chapter 5* and the statistical MI-5 model from *chapter 3*. Furthermore, we aimed to assess their value in general practice by comparing their predictive performance to the general practitioners' own prognosis in routine clinical practice. In 23 family practices in the Netherlands we recruited 203 patients with a new episode of shoulder pain who formed the Shoulder Pain Prediction (SPP) cohort. In this cohort we observed both the statistical and the expert-based models to perform suboptimal with AUC values of 0.580 and 0.625 respectively. However, we showed that this performance was only slightly better than the estimates of possible future outcome provide by the included 35 general practitioners (AUC value of 0.506). By this we highlighted the difficulties in estimating outcome in primary care patients with shoulder complaints. Our results suggest that prognostic models predicting the risk of persistent shoulder pain might be useful to primary care clinicians, but evaluated prognostic models showed inadequate external validity and would need to be improved.

In *chapter 7* we put the results of this thesis in a broader perspective, enumerate the main conclusions from this thesis, and provide recommendations for prognostic research. Briefly, researchers should always be aware of and address the disturbing influences of missing data and model stability when deriving a prediction model. Development of a predictive model is no straight forward procedure, in the process choices need to be made and the effects of these choices need to be evaluated and reported.