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

David Vergouw
METHODOLOGICAL ISSUES OF CLINICAL PREDICTION MODELS FOR SHOULDER PAIN IN GENERAL PRACTICE
This thesis was prepared within the EMGO+ Institute for Health and Care Research, Department of General Practice, VU University Medical Center Amsterdam, the Netherlands. The EMGO+ institute participates in the Netherlands School of Primary Care Research (CaRe), which was re-acknowledged in 2005 by the Royal Dutch Academy of Arts and Sciences (KNAW).

Financial support for the printing has kindly been provided by the VU University.

ISBN: 978-49-6191-606-8
Cover design and layout: David Vergouw
Printed by: Ipskamp drukkers - Enschede

© David Vergouw, Utrecht, the Netherlands, 2013.
All rights reserved. No parts of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system, without prior permission of the holder of the copyright.
METHODOLOGICAL ISSUES OF CLINICAL PREDICTION MODELS FOR SHOULDER PAIN IN GENERAL PRACTICE

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan de Vrije Universiteit Amsterdam, op gezag van de rector magnificus prof.dr. L.M. Bouter, in het openbaar te verdedigen ten overstaan van de promotiecommissie van de faculteit der Geneeskunde op maandag 11 februari 2013 om 13.45 uur in de aula van de universiteit, de Boelelaan 1105

door

David Vergouw

geboren te Utrecht
promotoren: prof.dr. H.E. van der Horst
            prof.dr.ir. H.C.W. de Vet

copromotoren: prof.dr. D.A.W.M. van der Windt
              dr. M.W. Heymans
CONTENTS

1. General introduction 07

2. Missing data and imputation 19
   a practical illustration in a prognostic study on low back pain

3. Imputation and stability 41
   the search for stable prognostic models in multiple imputed data sets

4. Logistic regression and CART analysis 67
   the search for stable prognostic models in multiple imputed data sets

5. Prediction in general practice 87
   comparing clinical consensus from a Delphi procedure with a statistical scoring system

6. Prediction models in general practice 117
   the value of prognostic models for persistent shoulder pain in general practice

7. General discussion 139

   Summary 159

   Samenvatting 166

   Dankwoord / Word of thanks 173
GENERAL INTRODUCTION
GENERAL INTRODUCTION

The main focus of this thesis is to study methodological issues in deriving a valid clinical prediction model and to assess its value in clinical practice. For this purpose we constructed multiple clinical prediction models for persistent musculoskeletal complaints, mainly shoulder pain by using various derivation methods. In this first chapter we will briefly introduce and define the main concepts used in this thesis. Subsequently our objectives are stated and an outline is provided of the chapters included in this thesis.
Prediction in musculoskeletal disorders

Musculoskeletal disorders are frequently occurring complaints which pose a major burden on health care and society.\(^1\)\(^2\) In the general Dutch population aged 25 years and over, the one year prevalence of regional musculoskeletal pain has been estimated at almost 75\%.\(^3\) Point prevalence estimates for regional musculoskeletal pain complaints (i.e., knee, hip, elbow, wrist, neck, shoulder and low back) range between 11 to 44\%\(^2\)-\(^6\) with low back and shoulder pain being the most commonly occurring complaints with a prevalence of around 44\% and 30\% respectively.\(^2\) Many people will have regional musculoskeletal pain in several sites\(^7\),\(^8\) and it is suggested that the seemingly different regional pain complaints share to a large extent similar risk and prognostic factors, and similar clinical course.\(^9\)

Although only 30\%-40\% of people reporting musculoskeletal pain consult a primary care physician for their complaints,\(^3\) musculoskeletal complaints have been shown to be the most frequently occurring reason for consulting a general practitioner (GP) in the Netherlands.\(^10\)

Managing musculoskeletal disorders in primary care poses difficulties. Complaints can be persistent or recurrent in many cases regardless of treatment.\(^3\) For instance, in patients presenting with shoulder complaints, symptoms will resolve after 6-12 months in only 40\%-70\% of all cases.\(^11\)-\(^13\)

Despite elaborate research into the pathology of shoulder pain, identifying the exact cause of musculoskeletal complaints in individual patients proves to be problematic, not only in general practice but also in clinical settings such as orthopaedics, rehabilitation medicine and pain clinics. This results in a high percentage of patients that cannot be provided with a specific
diagnosis. Since the early nineties musculoskeletal research has therefore focussed on exploring the determinants of an unfavourable course of musculoskeletal complaints, i.e., prognosis, rather than trying to find a precise cause. With these determinants, tools such as clinical prediction rules can be created that may complement the GP’s clinical judgement and expertise when distinguishing between complaints that will resolve spontaneously, complaints that can be managed in general practice, and complaints in need of referral for physiotherapy or other specialist care.

Clinical prediction models, often also referred to as risk scores, decision rules or prediction rules, have been developed to facilitate a quantitative estimate of the likely future outcome of e.g., low-back pain, knee pain, or shoulder pain.

**Clinical prediction models**

By using early disease symptoms and patient characteristics clinical prediction models generate absolute risks of particular outcomes of interest for individual patients. They can contribute to the estimation of the probability of having a specific condition (diagnosis) or the probability of a future outcome of a condition (prognosis). Physicians may subsequently use these estimates as assistive tools for making treatment decisions or for informing and advising patients. One of the most well known clinical prediction models is the APGAR score to quickly and summarily assess the health of new born children by evaluating the new born baby immediately after birth on five simple criteria (Appearance, Pulse, Grimace, Activity, Respiration).

From the increasing number of clinical prediction models published in the scientific literature it is clear that prediction is a popular topic. In prediction research three phases can be distinguished...
that need to precede the clinical application of a model. First a model needs to be derived from a relevant cohort study providing data on potential predictors and outcome. This is followed by a second phase of internal and external validation in which the model is tested in new cohorts of patients. And finally the third phase of developing a clinical prediction model is testing the impact of using the model in terms of patient outcomes and costs. These three phases are generally accepted as stages of model development, however in all phases i.e., model derivation, validation, and testing of impact, some methodological issues remain to be resolved.

**Issues in developing clinical prediction models**

One of the most difficult aspects of deriving a predictive model is finding a parsimonious set of predictors to form a simple yet good model that can consistently be applied in a broad patient population. This process can be hindered by multiple factors. Missing data is one of these factors. Absence of information in the derivation data set might affect the statistical power of the derived model or might bias the estimated associations between predictors and outcome. Although the advocated method of handling missing observations i.e., replacing missing observations with multiple estimations of plausible values (multiple imputation) is known to improve predictive modelling, it introduces two other troubling factors. Firstly, multiple imputation results in multiple imputed data sets. Combining these multiple data sets into a single predictive model is still a challenge. Additionally, multiple imputation introduces an extra source of instability i.e., imputation uncertainty, to the modelling process. This modelling process is already known to give unstable results due to the use of automated variable selection methods. This means that small differences in the data may lead to the identification of a different set of predictors. How to address
Chapter 1

This model stability especially in multiple imputed data sets is still an object of discussion.28

Next to giving unstable results, automated statistical predictor selection might overlook potential predictors obvious to clinicians. While the contribution of clinical expertise to the derivation of a prediction model is contribution of clinical expertise to the derivation of a prediction model is often minimal, a models’ clinical applicability might benefit from the incorporation of this knowledge. Gaining expert consensus on important predictors by for instance a Delphi procedure might therefore be a promising addition or alternative method for selecting important predictors.

Model derivation can be performed by using many different methods. Generally logistic regression (binary outcome) or Cox proportional hazard regression (time to event data) are used. However, alternative methods exist. One of the most frequently used alternative methods is recursive partitioning or decision tree building in the form of CART (Classification and regression Tree) analysis. According to proponents of this methodology, the distinct tree-like structure of these models makes them more clinically intuitive and therefore superior to traditional methods.30 However, decision trees have the tendency to depend strongly on the observed data and so give a falsely optimistic picture of the models’ true accuracy.31 Regression models on the other hand can also be overoptimistically fitted to the derivation data. It is therefore unclear whether decision trees or regression methods are better suited for deriving a valid and clinically applicable prediction model.

Overfitted models typically perform well in the derivation data, but modelled associations between predictors and outcome often do not apply as well as expected to new patients. Due to differences in patient characteristics or methodological deficiencies in model
derivation such as described above, the validity of a derived model may be poor.\textsuperscript{32} Although many scientific papers describe the development of a prediction model, a relatively small number describe the validation of a developed model. An even smaller number of studies tests the impact of a developed model in clinical practice, while model validity and clinical impact are most important for the implementation of a prediction model in clinical practice.

This thesis will address the above mentioned issues 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. Most chapters deal with predicting persistent shoulder pain in general practice, except for chapter 2 in which analyses using a data set on low back pain are described and discussed.

The topics addressed in this thesis can be categorized into the following objectives:

**Thesis objectives**

- To study the effects of missing information in baseline characteristics and outcome data on the derivation of a prognostic model for predicting persistent low back pain;
- To explore the stability of prognostic models derived using multiple imputed data sets;
- To study whether recursive partitioning can be used as an alternative for logistic regression in deriving a prognostic model for persistent shoulder pain;
- To identify whether recursive partitioning can be used as an alternative for logistic regression in deriving a prognostic model
for persistent shoulder pain;
- To identify the key clinical predictors of persistent shoulder pain;
- To assess the value of prognostic models in clinical practice by testing their external validity and comparison with GPs' own prognostic estimate.

Chapter 1

Thesis outline

The derivation of clinical prediction models in the presence of missing data is problematic. Chapter 2 aims to illustrate bias in derived models when using suboptimal methods for handling missing observations. A data set on low back pain provided the opportunity to study the effects of missing observations in either baseline characteristics (predictors) or outcomes at 6 months follow-up. This chapter also shows how multiple imputed data sets can be combined to form a single prognostic model.

Multiple imputation adopts multiple estimations of the values of missing observations to account for the uncertainty of each imputed value. As a result, multiple imputed data sets show small differences that introduce an extra source of model instability to the already unstable modelling process of automated variable selection. In chapter 3 a bootstrap resampling procedure was applied to a multiple imputed shoulder pain data set in order to explore the stability of a derived prediction model.

A promising alternative modelling technique that is said to result in more clinically intuitive models i.e., Classification and Regression Trees (CART), will be studied in chapter 4. Models predicting persistent shoulder pain intensity were constructed by logistic regression and CART. Since both methods are known to give overoptimistic results, internal validity was assessed for both models.
As clinical knowledge is expected to complement statistical modelling, *chapter 5* aims to identify the most important clinical predictors of persistent shoulder pain intensity. A Delphi study among 41 international health care professionals involved in the management of shoulder disorders was performed in order to reach consensus on important predictors. These predictors formed the basis of a prognostic model whose performance was compared to a statistically derived model.

How the derived models for persistent shoulder pain performed in new patients was studied in *chapter 6*. Next to external validation the value of the models in general practice was studied by comparing the models’ predictive performance with the prognostic estimate of the likely outcome of shoulder pain provided by the general practitioner.

The thesis concludes with a general discussion in *chapter 7* and a summary in both English and Dutch.
References


MISSING DATA AND IMPUTATION
a practical illustration in a prognostic study on low back pain

Published as:
ABSTRACT

Objectives: When designing prediction models by complete case analysis (CCA), missing information in either baseline (predictors) or outcomes may lead to biased results. Multiple imputation (MI) has been shown to be suitable for obtaining unbiased results. This study provides researchers with an empirical illustration of the use of MI in a dataset on low back pain (LBP), by comparing MI with the more commonly used CCA. Effects will be shown of imputing missing information on the composition and performance of prognostic models, distinguishing imputation of missing values in baseline characteristics and outcome data.

Methods: Data came from the Beliefs about Backpain (BeBack) cohort, a study of psychological obstacles to recovery in primary care back pain patients in the United Kingdom (UK). Candidate predictors included demographics, back pain characteristics and psychological variables. CCA was compared with MI within patients with complete outcome but missing baseline data (N=809) and patients with missing baseline or outcome data (N=1591). MI was performed by a Multiple Imputation by Chained Equations (MICE) procedure.

Results: Cases with missing outcome data (n=782, 49.1%) or with missing baseline data (n=116, 8%) both differed from complete cases regarding the distribution of some predictors and more often had a poor outcome. When comparing CCA with MI, model composition showed to be affected.

Conclusion: CCA can give biased results, even when only small amounts of data are missing. Now that MI is available in standard statistical software, we recommend it be used to handle missing data.
INTRODUCTION

Prognosis research, in particular the design of prediction models, has gained popularity over the past decade. These models enable the estimation of the probability of a specific outcome in individual patients. One of the main threats in the design of such models is missing information in either baseline (predictors) or follow-up outcomes. Once every effort has been made to ensure complete data collection, missing data need to be dealt with in the statistical analysis. For this a large number of methods exist, although in the maze of complicated techniques most researchers opt for the more simple approach of using available data only. Consequently, only subjects with complete information are used (i.e., complete case analysis, CCA). This method has proven to be appropriate when the excluded cases form a relatively small and representative portion of the entire dataset, or the values are missing completely at random (MCAR). However, this is seldom the case, as data is often missing for (un)observed reasons and therefore selectively missing. Then CCA is not only imprecise (because of loss of power), but may also give biased estimates of the investigated association(s).

A technique called imputation (i.e., replacing missing observation(s) with plausible estimates) is regarded as a good alternative for handling incomplete data. Various imputation methods exist, which mostly rely on the replacement of the missing value by another single value. Such simple methods assume MCAR independent of the percentage of missing data and may create more bias than they try to prevent. The advocated method for handling missing data is multiple imputation (MI). In contrast to a single value imputation, MI estimates the values of missing observations a number of times. This has several potential advantages; 1) it provides insight into the estimation variance and therefore accounts for imputation
uncertainty (this means that we cannot rely on one value being the true one, but there are more plausible values instead), 2) it prevents the regression estimates from being biased and standard errors from being too small, and thus 3) prevents hypothesis testing from being distorted. Therefore, MI is expected to result in more valid estimations of missing values than single imputation.

Although the interest in MI has grown in the last decade, the lack of use of this method coupled with the only very recent inclusion of it within regular statistical software (e.g., SPSS 17) has meant that the implications of the method are, as yet, not fully understood. This is especially relevant in the field of low back pain (LBP) where prognostic models have become increasingly popular, but where imputation techniques are hardly ever employed. Therefore, this study aims to provide researchers with an empirical illustration of the use of MI in a data set on LBP, by comparing MI with the more commonly used CCA. We will evaluate the effects of imputing missing information on the composition and performance of prognostic models derived in a cohort of primary care patients with LBP, distinguishing imputation of missing values in baseline characteristics and outcome data.

METHODS

The data set

For this empirical study we used data from the Beliefs about Backpain (BeBack) cohort, a study of psychological obstacles to recovery in patients with LBP consulting in primary care in the UK. This cohort consisted of 1591 adults who consulted their general practitioner with LBP between September 2004 and April 2006 in North Staffordshire and Central Cheshire in England. Patients were
excluded in case of red flag diagnoses (e.g., cauda equina syndrome, significant trauma, ankylosing spondylitis, cancers). Patients who consented to participation received a baseline questionnaire that included a wide range of predictors. Participants had to give permission for follow-up contact separately.

As in all prognostic research we had to deal with two sources of missing values 1) missing data on baseline characteristics (i.e., participants who did not fully complete the the baseline questionnaire) and 2) non-response at follow-up (i.e., participants who did not give written consent to be contacted at the follow-up points or who did not fully complete the follow-up questionnaire).

Ethical approval for the study was obtained from the North Staffordshire and Central Cheshire Research Ethics Committees.

**Measures**

For the purpose of this prognostic study the intensity of LBP, a secondary outcome in the BeBack study, was used as outcome. Pain intensity, rated on a 0-10 numerical scale, was dichotomized in order to perform logistic regression: patients who improved 30% or more in pain intensity were considered improved by a clinically important degree and those who improved less than 30% were denoted as having persistent symptoms (i.e., not improved by a clinically important degree). From the available follow-up time points (3, 6 and 12 months), we used the outcome assessed 6 months after initial consultation. Candidate predictors included: 1) socio-demographic variables (gender, age and employment status), 2) back pain characteristics (disability Roland and Morris Disability Questionnaire (RMDQ)), duration of complaint, patients’ recall of the last pain-free month, pain intensity (0-10 numeric rating scale), co-morbidity (yes/no questions), patients’ perceptions about their back problem (modified revised illness perceptions questionnaire
chapter 2

(IPQR⁸,¹⁴), 3) other psychological obstacles to recovery (anxiety and depression, self-efficacy, fear of movement, and catastrophizing)⁹ and physical activities (before onset of the current episode and limitations due to the low back problem). For all candidate predictors the linearity assumption was checked. When the relationships between variables and outcome did not resemble linearity, variables were categorized (3 categories) or dichotomized. Furthermore, the variables with a univariable association with the outcome were checked for co-linearity using Pearson’s r. In case of co-linearity (r ≥0.7) the variables which we thought could most easily be obtained in clinical practice were retained.

Missing data

Missing data were handled using two different methods. The advocated method in case of missing data MI, was compared to the more commonly used but suboptimal CCA. To perform MI several techniques are available under several software packages. In this study missing values were filled by estimating their values using Multiple Imputation by Chained Equations (MICE) procedure.¹⁵ We used all observed patient data including outcome data on pain intensity to estimate imputation values in five replicate imputed datasets.⁴ This method was applied for imputing the two different sources of missing data (missing values in baseline and outcome data). First, in line with most prognostic studies, we included only participants with complete follow-up data and imputed missing baseline characteristics. However, by doing so a substantial portion of gathered data is still left unused. Therefore, secondly, all participants with incomplete or completely missing follow-up data were used to impute missing values on both outcomes and baseline characteristics. The latter analysis requires imputation of a much larger proportion of the data, but retains a population that is more
similar to the source population.

**Statistical analysis**

The derived prognostic models were all based on multivariable logistic regression with a backward selection procedure (critical value of $\alpha=.05$). To check whether the analysis was sufficiently powered to build a multivariable prognostic model the number of events per variable (EPV) was calculated.$^{16,17}$

**Variable selection**

For both methods CCA and MI, the first step in variable selection was performed by a univariable ($p\leq .05$) analysis on each of the imputed datasets. Next a multivariable regression with a backward selection procedure was applied including only univariably associated predictors. Since MI produced five imputed datasets, the backward selection procedure needed to be applied to each of these replicate data sets. From the resulting five models, predictors which appeared in at least two of the five models (an inclusion frequency of $\geq 40\%$)$^{18}$ were selected to form the overall model. This overall model was then applied to each of the five imputed data sets and the final model was formed by averaging the regression coefficient estimates. Whether all included predictors significantly contributed to the final model was tested using a likelihood ratio test$^{19}$ with a critical $p$-value of $p=.05$. In case of a non-significant likelihood ratio test, the predictor was eliminated from the final model.

**Model evaluation**

To evaluate the influence of different methods of dealing with missing data on the derived models, primarily the model composition (combination of predictors) was considered. In addition,
chapter 2

the model performance parameters discrimination, calibration and explained variance were assessed.

Discrimination, i.e., how well a model distinguishes between patients with and without persistent symptoms, is quantified by the c-index that, for binary outcomes, is identical to the area under the receiver operating curve (AUC).\textsuperscript{20}

Calibration indicates the agreement between predicted and observed probabilities and was measured by computing the slope of the calibration plot.

Nagelkerke’s $R^2$ was used as a measure of the explained variance.

Internal validation

Since the apparent performance of a predictive model is typically better in the study population compared to the performance in the source population,\textsuperscript{21} the derived models’ performance likely overestimates the models’ true performance. Whether the regression coefficients were optimistically estimated was analyzed by a 200 sample bootstrap procedure for internal validation.\textsuperscript{22} Subsequently the c-indices for the overoptimistic models were adjusted.

Software

All analyses were performed using the R-statistics software (version 2.4.0). This includes the R Design package that was used for the CCA and the MICE package which was used for the MI.

RESULTS

Study population

Figure 1 shows the inclusion and follow-up of participants in the
BeBack study. 1591 adults who had consulted their GP with LBP consented to participate in the baseline data collection. Baseline characteristics are shown in Table 1. Of the 1591 baseline participants, 302 (19%) did not give written consent for further contact, leaving 1289 participants. Of these 1254 were mailed the 6 month questionnaire. At 6 months follow-up, 810 out of 1254 (64.6%) participants actually completed the questionnaire, resulting in a total response rate of 50.9% (810/1591).

**Figure 1. Flowchart of patient participation and data collection**
Table 1. Patient characteristics from all available data at baseline (n=1591) and from the different missing Value groups.

<table>
<thead>
<tr>
<th>variable</th>
<th>baseline (n=1591)</th>
<th>complete cases (n=693)</th>
<th>participants with missing baseline data (n=116)</th>
<th>participants with missing follow-up data (n=782)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean (SD)</td>
<td>mean (SD)</td>
<td>mean (SD)</td>
<td>mean (SD)</td>
</tr>
<tr>
<td>Demographic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age (years)</td>
<td>43 (10.3)</td>
<td>46 (9.9)</td>
<td>45 (8.6)</td>
<td>42 (10.6)</td>
</tr>
<tr>
<td>gender (male); n(%)</td>
<td>661 (41.5)</td>
<td>279 (40.3)</td>
<td>36 (31.0)</td>
<td>346 (44.2)</td>
</tr>
<tr>
<td>currently employed; n(%)</td>
<td>1177 (74.0)</td>
<td>529 (76.3)</td>
<td>89 (76.4)</td>
<td>576 (73.7)</td>
</tr>
<tr>
<td>Disease characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>upper extremity pain past 4 weeks; n(%)</td>
<td>713 (44.8)</td>
<td>323 (46.6)</td>
<td>58 (50.0)</td>
<td>341 (43.6)</td>
</tr>
<tr>
<td>head pain past 4 weeks; n()</td>
<td>341 (21.4)</td>
<td>154 (22.2)</td>
<td>27 (23.6)</td>
<td>164 (21.0)</td>
</tr>
<tr>
<td>usual LBP level last 2 weeks [0-10]</td>
<td>4.6 (2.7)</td>
<td>4.8 (2.5)</td>
<td>4.76 (2.9)</td>
<td>4.5 (2.79)</td>
</tr>
<tr>
<td>duration of the current episode; n(%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤3 months</td>
<td>1032 (64.3)</td>
<td>471 (68.0)</td>
<td>56 (60.9)</td>
<td>496 (66.6)</td>
</tr>
<tr>
<td>4-6 months</td>
<td>148 (9.3)</td>
<td>68 (9.8)</td>
<td>9 (9.8)</td>
<td>71 (9.5)</td>
</tr>
<tr>
<td>&gt;6 months</td>
<td>583 (22.6)</td>
<td>154 (22.2)</td>
<td>27 (29.3)</td>
<td>178 (23.9)</td>
</tr>
<tr>
<td>disability level at baseline; n(%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no-little [0.0-7.9]</td>
<td>859 (54.0)</td>
<td>362 (52.2)</td>
<td>63 (54.3)</td>
<td>434 (55.6)</td>
</tr>
<tr>
<td>moderate [8.0-10.9]</td>
<td>148 (9.3)</td>
<td>75 (10.8)</td>
<td>13 (11.2)</td>
<td>60 (7.7)</td>
</tr>
<tr>
<td>bad [11.0-24.0]</td>
<td>583 (36.6)</td>
<td>256 (36.9)</td>
<td>40 (34.5)</td>
<td>287 (36.7)</td>
</tr>
<tr>
<td>used medication w/o prescription; n(%)</td>
<td>756 (47.5)</td>
<td>340 (49.1)</td>
<td>44 (37.9)</td>
<td>372 (47.6)</td>
</tr>
<tr>
<td>consulted any form of healthcare</td>
<td>1427 (89.7)</td>
<td>651 (93.9)</td>
<td>100 (86.2)</td>
<td>676 (86.4)</td>
</tr>
<tr>
<td>Psychological factors</td>
<td>8.3</td>
<td>(4.5)</td>
<td>8.2</td>
<td>(4.4)</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-----</td>
<td>-------</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>anxiety [0-21]</td>
<td>6.5</td>
<td>(4.4)</td>
<td>6.3</td>
<td>(4.1)</td>
</tr>
<tr>
<td>depression [0-21]</td>
<td>9.9</td>
<td>(7.9)</td>
<td>9.5</td>
<td>(7.5)</td>
</tr>
<tr>
<td>catastrophizing [0-36]</td>
<td>15.5</td>
<td>(8.2)</td>
<td>15.5</td>
<td>(8.2)</td>
</tr>
<tr>
<td>diversion [0-36]</td>
<td>16.7</td>
<td>(5.2)</td>
<td>16.7</td>
<td>(5.2)</td>
</tr>
<tr>
<td>IPQR – consequences [0-30]</td>
<td>20.5</td>
<td>(3.8)</td>
<td>21.0</td>
<td>(3.8)</td>
</tr>
<tr>
<td>IPQR – personal controle [0-30]</td>
<td>17.0</td>
<td>(3.4)</td>
<td>17.2</td>
<td>(3.3)</td>
</tr>
<tr>
<td>IPQR – timeline [0-30]</td>
<td>19.6</td>
<td>(5.8)</td>
<td>19.9</td>
<td>(5.8)</td>
</tr>
<tr>
<td>IPQR – symptoms original [0-15]</td>
<td>11.9</td>
<td>(4.1)</td>
<td>11.7</td>
<td>(4.0)</td>
</tr>
<tr>
<td>IPQR – psychological attribution [0-30]</td>
<td>15.0</td>
<td>(4.1)</td>
<td>15.0</td>
<td>(4.0)</td>
</tr>
<tr>
<td>IPQR – immunity [0-15]</td>
<td>5.3</td>
<td>(1.9)</td>
<td>5.2</td>
<td>(1.9)</td>
</tr>
<tr>
<td>IPQR – accident / chance [0-10]</td>
<td>6.0</td>
<td>(1.9)</td>
<td>6.0</td>
<td>(1.9)</td>
</tr>
<tr>
<td>self efficacy score [0-60]</td>
<td>37.8</td>
<td>(14.6)</td>
<td>38.4</td>
<td>(14.0)</td>
</tr>
</tbody>
</table>

**IPQR** modified revised Illness Perceptions Questionnaire
Imputation of missing baseline characteristics

From the 810 participants who completed the 6 months questionnaire, 809 participants had complete follow-up data (one participant had a missing value on the outcome LBP intensity) and 116 (14.3%) had missing values on some of the recorded baseline characteristics. Thus the CCA sample consisted of 693 participants. As can be seen in Table 1, compared to the complete case sample, among the 116 participants with missing baseline characteristics, there were fewer males (31% vs. 40%), they reported fewer health care consultations during the four weeks prior to baseline (86% vs. 94%) and used fewer over the counter medication prior to baseline (38% vs. 49%). Participants with missing baseline characteristics therefore did not form a representative portion of the participants with complete outcome data and imputation was desired. When missing values were estimated by MI, an imputed dataset (MI-809) was created. In this dataset, no changes in the distributions of individual baseline characteristics were observed (data not shown).

Imputation of missing values on outcome

The 782 participants (51.2%) with no recorded 6 months follow-up data differed from 809 participants with complete follow-up data. Participants without follow-up data were more often men (44% versus 39%), were on average younger (42.2 vs. 45.5 years), and reported fewer health care consultations during the four weeks prior to baseline (87% vs. 93%). Other baseline characteristics like employment status, symptom duration, disability and each of the IPQR dimensions were almost identical for participants with and without 6 months follow-up data. Because of the differences between participants with and without complete follow-up data, there was a selective but non-informative loss of outcome data and therefore
imputation was desired. When missing values on outcome were accounted for by MI, the MI-1591 dataset was created. When compared to the complete cases, the distribution of the outcome measure persistent LBP intensity changed in the MI-1591 dataset (Table 2). 54.3% of the complete cases were denoted as being recovered, where in the MI-1591 dataset only 46.9% was denoted as being recovered. Thus, a large part (57.9%) of all the participants with missing follow-up data was denoted by MI as having persistent pain at follow-up.

Table 2. Distribution of the outcome measure persistent LBP in subjects with complete data and the imputed data sets (averaged over five data sets after multiple imputation).

<table>
<thead>
<tr>
<th>Persistent LBP intensity</th>
<th>complete cases (n=693)</th>
<th>MI–809 (n=809)</th>
<th>MI–1591 (n=1591)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases (persistent)</td>
<td>317 (45.7)</td>
<td>392 (48.4)</td>
<td>845 (53.1)</td>
</tr>
<tr>
<td>non–cases (recovered)</td>
<td>376 (54.3)</td>
<td>417 (51.5)</td>
<td>746 (46.9)</td>
</tr>
</tbody>
</table>

MI-1591 all data including baseline characteristics from subjects lost to follow-up
MI-809 all data minus loss to follow-up

Effects of imputation on model composition

Table 3 shows the three derived prognostic models: CCA based on complete cases only, MI-809 based on participants who completed the 6 months follow-up but using imputation for missing values on baseline characteristics, and MI-1591 based on imputation of missing baseline characteristics and missing outcome data. All models were sufficiently powered (EPV>10) since 26 variables showed a univariable association with outcome (data not presented), resulting in an EPV of 13.5. In the five imputation datasets the EPV
varied between 32.1 and 32.9. Co-linearity did not occur. The largest models were the CCA and MI-809 models, both containing 9 predictors. The MI-1591 model contained 6 predictors. For predictors occurring in more than one model, the directions of the associations (i.e., regression coefficients) between the same selected predictors and the outcome were the same in all models (Table3).

**Effects of imputation on model composition**

Table 3 shows the three derived prognostic models: CCA based on complete cases only, MI-809 based on participants who completed the 6 months follow-up but using imputation for missing values on baseline characteristics, and MI-1591 based on imputation of missing baseline characteristics and missing outcome data. All models were sufficiently powered (EPV>10) since 26 variables showed a univariable association with outcome (data not presented), resulting in an EPV of 13.5. In the five imputation datasets the EPV varied between 32.1 and 32.9. Co-linearity did not occur. The largest models were the CCA and MI-809 models, both containing 9 predictors. The MI-1591 model contained 6 predictors. For predictors occurring in more than one model, the directions of the associations (i.e., regression coefficients) between the same selected predictors and the outcome were the same in all models (Table3).

Complete 6 months follow-up based models (CCA and MI-809), differed due to the in- or exclusion of a small portion (8%) of participants with missing baseline values. Although both models included the same number of predictors, they did not consist of exactly the same predictors. The predictor “catastrophizing” from the MI-809 model was replaced by the predictor “over the counter medication” in the CCA model and therefore the regression coefficient estimates and ranking of the predictors based on the strength of their associations with outcome differed between models.
When missing values on outcome were included in the imputation routine (MI-1591), baseline characteristics from a large portion of participants (51.2%) were added to the dataset. Accordingly larger differences in model composition were observed. Some of the predictors that were denoted as having stronger associations with persistent LBP in participants with complete outcome data (i.e., bothersomeness, health service use and duration of the current LBP episode) were not included in the MI-1591 model. As a result the number of included predictors was smaller in the MI-1591 model.

**Model performance**

Parameters for model performance are presented in Table 4. Calibration, explained variance and the apparent discrimination were the highest for the MI-809 model. When missing values on outcome were included in the imputation routine (MI-1591), these numbers were lower. However, according to the internal validation test, this model was the least overoptimistic. After correction for overoptimism the results for discrimination were still higher for the MI-809 model (adjusted c-index 0.756) and lower for the MI-1591 model (adjusted c-index 0.721).

**Table 4. Performance measures for each prognostic model**

<table>
<thead>
<tr>
<th></th>
<th>MI–809</th>
<th>MI–1591</th>
<th>CCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>calibration slope</td>
<td>1.005</td>
<td>1.033</td>
<td>1.024</td>
</tr>
<tr>
<td>R²N</td>
<td>0.272</td>
<td>0.201</td>
<td>0.248</td>
</tr>
<tr>
<td>Ac (95% CI)</td>
<td>0.768 (.736–.800)</td>
<td>0.724 (.700–.749)</td>
<td>0.759 (.724–.793)</td>
</tr>
<tr>
<td>Opt</td>
<td>0.012</td>
<td>0.003</td>
<td>0.014</td>
</tr>
<tr>
<td>Oc</td>
<td>0.756</td>
<td>0.721</td>
<td>0.744</td>
</tr>
</tbody>
</table>

Ac = apparent c-index  
CCA = complete case analysis  
MI = multiple imputation using 5 imputation files  
Oc = optimism corrected c-index  
Opt = estimation of the overoptimism  
R²N = explained variance (Nagelkerke’s R-squared)  
95% CI = 95% confidence interval
DISCUSSION

In prognostic research it is common practice to exclude participants with incomplete baseline or follow-up data. However, not using the full study sample but only participants with complete data may reduce the models’ validity. In this study the composition of prognostic models based on complete cases was compared with those based on multiple imputation. We empirically illustrated that included predictors and regression coefficient estimates alter when parts of the full study sample were used. Even missing a small non-representative portion of the entire dataset (8% missings in the baseline characteristics) led to variation in model composition. With increasing numbers of missing data, this effect became more evident: even variables identified as strong predictors by complete case analysis were not included in the full study sample based model.

In the current study we illustrated the effects of different methods of handling missing data on model composition, by comparing CCA with MI for handling missing values in prognostic models. The reason for which values are missing, i.e., the underlying “cause” or missing data mechanism, determines whether a certain method of analysis provides reliable and unbiased results. CCA for instance is known to only be appropriate in the rare instances in which data is missing completely at random (MCAR) or the excluded cases form a relatively small and representative portion of the entire dataset. In all other situations, CCA is regarded as a suboptimal method of analysis.

Since the assumptions for valid CCA results did not hold in our study sample (i.e., participants with missing values did not form a representative portion of the entire dataset), our CCA results are expected to be biased. In our illustration missing values appeared to
be related to other observed information. This type of missing data is called Missing At Random (MAR).\textsuperscript{5} In this more frequently occurring type of missing data, MI has in previous simulation studies been shown to be a suitable method to obtain results similar as those which would be obtained when the dataset would be complete.\textsuperscript{2,23} Furthermore, even when the MAR assumption has been violated MI methods have been shown to be always at least as good as CCA.\textsuperscript{23} Therefore we expect our MI models to be superior to our CCA models, and interpret difference between the models as bias for the CCA.

Although there are many statistical software packages that support MI, comparing those was beyond the scope of this study. Here we were interested in illustrating the advantages of MI over CCA for a clinical audience.

Although, it has been recognised for quite some time that excluding participants with missing data causes bias, CCA remains a popular technique in prognostic research. There are a number of reasons for this. Firstly, CCA was until recently the norm in standard statistical software, and performing MI was far more complicated. Now with the recent introduction of MI in standard statistical software (e.g., SPSS 17), performing MI has become easily accessible. Secondly, imputation is sometimes incorrectly believed to be only necessary when large portions of data are missing. However, CCA can give biased results, even when only small amounts of data are missing,\textsuperscript{24} as we showed in our example with missing baseline values. Thirdly, intuitively researchers may be reluctant to fill in data that they have not observed. Finally, in our study there is an ethical issue of whether it is justified to impute outcome data for individuals who did not give consent for further participation. However, since only information that was consented for was used to impute missing follow-up data, we argue that this approach is not unethical. An
imputed dataset reflects more closely the target population for which the prognostic model is intended, making better use of all data collected in consenting participants and thus resulting in more valid outcomes.

The predictors included in the model were quite different, when using the different methods of handling missing data. This may have consequences for clinical practice. As clinicians think in terms of patient (risk) profiles when identifying patients with a poor prognosis, it makes a lot of difference when the variable ‘over the counter medication’ replaces ‘CSQ catastrophizing’, as was the case when the model obtained by CCA was compared to the MI-809 model. For clinicians these are different patient groups, which may require a different type of intervention. From a statisticians point of view, the ability to correctly predict the outcomes would be most interesting. Note that the parameters for model performance did not play a decisive role in the choice of the best imputation model, we only presented these for information.

**Limitations**

Previous simulation studies, which evaluated the impact of MAR missing data, have identified MI as being superior of identifying the “true” predefined predictors of outcome compared to CCA. An advantage of our study sample is that it resembled a dataset used in clinical prognostic research with respect to the strength of associations between predictors and outcome, and with respect to the percentage of missing values. On the other side, it may be a limitation that in our study knowledge of the “true” predictors of outcome is lacking because of the clinical approach. However as simulation studies already showed that MI is superior to CCA methods, we found describing the differences between the two
Another limitation is that by using the MI approach we assumed the missings to be missing at random. However, there is a third type of missing data, called missing not at random (MNAR), when the reason for a missing value depends on information not available in the dataset which makes the missing data unpredictable and MI useless. It is impossible to determine from the obtained study data whether missing data is MAR or MNAR; this can only be reasoned or speculated.\textsuperscript{25} In our study there was a high percentage of missings on the outcome variable which could be MAR or MNAR. Missing outcomes in more than half of the population is not uncommon in prognostic studies with a long follow-up period. In our study this missing data in outcome was caused by lack of consent for follow-up measurements. As this lack of consent occurred at baseline, it is plausible to assume that these missings are MAR instead of MNAR.

**Conclusions**

In conclusion, in a large data set on the clinical course of low back pain the effect of two sources of missing data (i.e., a small number of missing values in baseline characteristics and a large number of missing values in follow-up data) could be clearly illustrated. Handling of missing data by MI showed that model composition was affected by missing data, and thus CCA may lead to a biased prediction model. Since in clinical (and simulated) data MI is known to give less biased results in at least MAR data, using MI reduces the risk of biased results. Because of this bias and of its possible clinical consequences, we recommend that MI is used in all studies with missing data.
References


IMPUTATION AND STABILITY
the search for stable prognostic models in multiple imputed data sets

Published as:
ABSTRACT

Objectives: In prognostic studies model instability and missing data can be troubling factors. Proposed methods for handling these situations are bootstrapping (B) and Multiple imputation (MI). The authors examined the influence of these methods on model composition.

Methods: Models were constructed using a cohort of 587 patients consulting between January 2001 and January 2003 with a shoulder problem in general practice in the Netherlands (the Dutch Shoulder Study). Outcome measures were persistent shoulder disability and persistent shoulder pain. Potential predictors included socio-demographic variables, characteristics of the pain problem, physical activity and psychosocial factors. Model composition and performance (calibration and discrimination) were assessed for models using a complete case analysis, MI, bootstrapping or both MI and bootstrapping.

Results: Results showed that model composition varied between models as a result of how missing data was handled and that bootstrapping provided additional information on the stability of the selected prognostic model.

Conclusion: In prognostic modelling missing data needs to be handled by MI and bootstrap model selection is advised in order to provide information on model stability.
INTRODUCTION

In healthcare predicting how long it takes for an episode of musculoskeletal pain to resolve can be difficult. Outcome varies between patients and over time. Although clinicians can be relatively good “prognosticians”\(^1,2\) clinical judgment and intuition can be incorrect and difficult to quantify or to be made explicit. To understand the ingredients that contribute to correct prognosis and to improve upon clinical judgment, clinical prediction rules can be useful. These provide a quantitative estimate of the absolute risk of particular outcomes of interest for individual patients, which may subsequently be used to support decisions regarding treatment. Until now, several clinical prediction rules have been developed in the field of musculoskeletal pain, for example to estimate the outcome of low back,\(^2-4\) knee,\(^5\) or shoulder\(^6\) pain.

In the development of clinical prediction models, researchers frequently use a regression analysis with a backward or forward selection strategy. However, this methodology may result in overoptimistically estimated regression coefficients, omission of important predictors and random selection of less important predictors. As a result derived models may be unstable.\(^7\) Incorporating a bootstrap resampling procedure in model development has been suggested to provide information on model stability.\(^8-11\) Since bootstrapping mimics the sampling variation in the population from which the sample was drawn it is expected to produce a model which better represents the underlying population.\(^9-11\)

Another problem occurring in prognostic studies is missing data. Multiple imputation (MI), which uses all observed information, was shown to be superior to other imputation techniques like single regression imputation.\(^12,13\) Though, MI is not yet frequently used in
predictive modelling and model stability is hardly ever accounted for in MI approaches. It has been shown that for low back pain extending MI with a bootstrapping procedure provides an accurate model selection and information on model stability.\textsuperscript{14} However, generalizability of this method was never tested in other patient datasets.

Therefore, the objective of our research was to examine the influence of bootstrapping and multiple imputation on model composition and stability in a shoulder pain data set with missing values.

**METHODS**

**Study population**

We used data from the Dutch Shoulder Study (DSS).\textsuperscript{15} This cohort consists of 587 patients who consulted their general practitioner (GP) with a new episode of shoulder disorders. Inclusion criteria were: no GP consultation or treatment received for the afflicted shoulder in the preceding three months. Exclusion criteria were: dementia, severe psychiatric or physical conditions (i.e., fractures or dislocation in the shoulder region, rheumatic diseases, neoplasms, neurological or vascular disorders). The ethics review board of the VU University medical centre approved the study protocol.

**Outcome measures**

We focused on two outcome measures; persistent shoulder disability (16-item SDQ; 0-100)\textsuperscript{16} and persistent shoulder pain intensity (Numeric Rating Scale; 0-10)\textsuperscript{17}. To define ‘persistence’ baseline scores were subtracted from follow-up scores. An optimal cut-off point was defined by studying the relationship between the change
scores and a secondary outcome measure ‘patient perceived recovery’. Patients were denoted as recovered when they characterized their complaints as ‘fully recovered’ or ‘very much improved’.

By constructing Receiver Operating Characteristic (ROC) curves with patient perceived recovery as the external criterion, the optimal cut off point (i.e., that point that yields the lowest overall misclassification) was determined. According to this analysis a 50% decrease in disability and pain intensity compared to baseline was considered a minimal important change, and was used as a cut-off value to dichotomize both outcome measures. Patients who improved less than 50% were denoted as having persistent pain or disability. Outcomes were measured three months after enrolment by postal questionnaire.

**Prognostic factors**

Based on a systematic review of the literature a set of candidate predictors was selected, including; demographic variables, characteristics of the shoulder pain problem, physical and psychological factors (see Table 1). The following questionnaires were used to gather information on psychological factors: the Pain Coping and Cognition List (PCCL: pain coping, catastrophizing, internal and external locus of control), the 4 Dimensional Symptom Questionnaire (4DSQ: anxiety, depression, somatisation, distress), the Fear-Avoidance Beliefs Questionnaire (FABQ: fear-avoidance) and the Tampa Scale for Kinesiophobia (TSK: kinesiophobia). Within 10 days after consulting the GP, participants completed a baseline questionnaire to assess potential predictors.
Analysis

For all continuous predictors the linearity assumption was checked. When the relationships between variables and outcome did not resemble linearity, variables were categorized (3 categories) or dichotomized. Although this causes loss of information,\textsuperscript{25} these procedures were retained since they are part of the frequently used standard statistical methodology in predictive modelling. Variables were checked for (multi)collinearity using Pearson’s r, given that correlated variables can disturb variable selection in multivariable regression.\textsuperscript{26} In case of correlation ($r \geq 0.5$) the variables which could most easily be obtained in clinical practice by the physician were retained.

To reduce the initial number of variables, an univariable analysis ($\alpha > 0.157$) was performed in both the imputed and unimputed data sets, thus all analyses were preceded by this pre-selection. The subsequent analyses were all based on a multivariable analysis with a backward selection strategy and a stopping rule of $\alpha = 0.157$. This significance level is available in many statistical software packages and results have been shown to be comparable with the more complex Akaike Information Criterion (AIC).\textsuperscript{27} The number of events per variable (EPV) was calculated for each method to check whether the analysis was sufficiently powered (EPV $> 10$).\textsuperscript{28} The checklist proposed by Harrell\textsuperscript{29} for multivariable modelling was followed where possible. To study the effect of missing data and model stability on model composition, the following four methods were compared:

1) **Complete Case Analysis (CCA)**

To handle missing data, subjects with missing values on any of the variables were omitted and only those subjects with information on all variables in the model were included for analysis.
2) **Multiple imputation (MI-5)**

Missing values were imputed using a Multivariate Imputation by Chained Equations (MICE) procedure with the “predictive mean matching” as imputation method. All available data including outcome measure were used in the imputation method. We generated five imputed data sets (MI-5). Multivariable regression was applied to each of the 5 imputed data sets. From these 5 models, predictors which appeared in at least 2 models (a Inclusion Fraction of ≥40%) qualified for the final model. Whether these predictors significantly contributed to the final model was tested using a likelihood ratio test with a critical p-value of p = .157. Predictors were dropped from the final model in case of a nonsignificant (P>.157) likelihood ratio.

3) **Bootstrapping (B)**

A two-step bootstrap model selection procedure was applied to provide information on model stability. First 500 samples with replacement were taken from the complete case data set. In each sample a multivariable model was built. To be consistent with the MI-5 method, predictors which appeared in ≥40% of these models qualified for the second step. In this second step 500 new complete case samples were taken and in each of which a multivariable model was built using the predictors from the first step. These 500 models provided information on model stability (i.e., which combination of predictors is most frequently selected in the model).

4) **Multiple imputation + bootstrapping (MI-5+B)**

Missing data was imputed using the MICE procedure and five imputed data sets were created. In each of the five imputed data sets the two step bootstrap model selection procedure as described above was applied. Information on model stability was provided by
studying which combination of predictors occurred most frequently in 2500 data sets.

**Internal validation**

The apparent performance of a predictive model is typically better in the data set in which the model has been developed compared to its performance in another similar data set.\(^{32}\) This phenomenon is called overoptimism. Using a n=200 samples bootstrap procedure for internal validation\(^ {33}\) the performance of each developed model was tested in similar populations as in the derivation sample. This method was used to estimate the overoptimism of the derived models, and to adjust the measures of performance.

**Model evaluation**

Derived models were evaluated by comparing the model’s composition (combination of predictors).

Next several measures of predictive performance were considered. Discrimination refers to how well a model distinguishes between patients with and without persistent symptoms and is quantified by the c-index that, for binary outcomes, is identical to the area under the ROC curve (AUC).\(^ {34}\) The c-index varies between 0.5 and 1, with 0.5 indicating no discrimination above chance and 1 indicating perfect discrimination. The agreement between predicted probabilities and observed probabilities is called calibration and was measured by computing the slope of the calibration plot (predicted probabilities against observed frequencies). Well-calibrated models have a slope of 1. As a measure of the explained variance Nagelkerke’s R\(^2\) was computed.

**Software**

All analyses were performed using the R-statistics software (version
2.4.0). The R Design package was used for the CCA, MICE was used for the MI and additional routines were developed for applying the bootstrap.

RESULTS

The baseline patient characteristics are listed in Table 1. After three months 517 patients (88%) returned the follow-up questionnaire. Subjects lost to follow-up were younger (mean difference of 7 years) and showed more often an acute onset (47% versus 37%). Due to non-response the percentage of missing data was largest for the outcome measures (shoulder disability 12.3% and shoulder pain intensity 12.9%). Other (baseline) variables had missing values within the range of 0 to 9.2%. The combination of missing values in CCA resulted in the exclusion of 24.7% (disability model) and 28.8% (pain intensity model) of participants.

In the CCA 12 variables showed a univariable association with persistent disability and 16 with persistent pain, resulting in an EPV of 11.9 for pain intensity and 17 for shoulder disability. In the five imputation data sets the EPV varied between 19.1 and 19.6 for disability and between 13.5 and 13.8 for pain intensity. This means that the analyses were sufficiently powered (with a sufficient number of cases in the models) to reliably estimate the associations between predictors and outcome.

Model composition

For all presented models, the directions of the associations (i.e., regression coefficients) between the selected predictors and outcome were the same for both disability and pain (data not presented).
<table>
<thead>
<tr>
<th>variable</th>
<th>n</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age (years); mean (SD)</td>
<td>51</td>
<td>(14)</td>
</tr>
<tr>
<td>gender (male); n(%)</td>
<td>292</td>
<td>(50)</td>
</tr>
<tr>
<td>education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>210</td>
<td>(36)</td>
</tr>
<tr>
<td>middle</td>
<td>234</td>
<td>(40)</td>
</tr>
<tr>
<td>high</td>
<td>135</td>
<td>(23)</td>
</tr>
<tr>
<td><strong>Disease characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shoulder complaints in the past year</td>
<td>321</td>
<td>(55)</td>
</tr>
<tr>
<td>neck complaints in the past year</td>
<td>252</td>
<td>(43)</td>
</tr>
<tr>
<td>duration of complaints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-6 weeks</td>
<td>205</td>
<td>(35)</td>
</tr>
<tr>
<td>7-12 weeks</td>
<td>139</td>
<td>(24)</td>
</tr>
<tr>
<td>&gt;3 months</td>
<td>242</td>
<td>(41)</td>
</tr>
<tr>
<td>gradual onset (vs. Acute)</td>
<td>363</td>
<td>(62)</td>
</tr>
<tr>
<td>shoulder pain [0-10]; mean (SD)</td>
<td>4.8</td>
<td>(2)</td>
</tr>
<tr>
<td>shoulder disability [0-100]; mean (SD)</td>
<td>59.9</td>
<td>(24)</td>
</tr>
<tr>
<td>both shoulders afflicted</td>
<td>74</td>
<td>(13)</td>
</tr>
<tr>
<td><strong>co-morbidity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>upper extremity joint pain</td>
<td>245</td>
<td>(42)</td>
</tr>
<tr>
<td>neck pain</td>
<td>197</td>
<td>(34)</td>
</tr>
<tr>
<td>lower extremity joint pain</td>
<td>174</td>
<td>(30)</td>
</tr>
<tr>
<td>low back pain</td>
<td>139</td>
<td>(24)</td>
</tr>
<tr>
<td>high back pain</td>
<td>53</td>
<td>(9)</td>
</tr>
<tr>
<td><strong>Psychological factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>psychological complaints</td>
<td>55</td>
<td>(9)</td>
</tr>
<tr>
<td>pain coping [0-6]; mean (SD)</td>
<td>2.98</td>
<td>(0.98)</td>
</tr>
<tr>
<td>catastrophizing [0-6]; mean (SD)</td>
<td>2.2</td>
<td>(0.8)</td>
</tr>
<tr>
<td>internal locus of control [0-6]; mean (SD)</td>
<td>3.3</td>
<td>(0.9)</td>
</tr>
<tr>
<td>external locus of control [0-6]; mean (SD)</td>
<td>3.2</td>
<td>(0.88)</td>
</tr>
<tr>
<td>anxiety [0-24]; mean (SD)</td>
<td>0.3</td>
<td>(1.2)</td>
</tr>
<tr>
<td>depression [0-12]; mean (SD)</td>
<td>0.2</td>
<td>(1.3)</td>
</tr>
<tr>
<td>somatisation [0-32]; mean (SD)</td>
<td>3.3</td>
<td>(4.1)</td>
</tr>
<tr>
<td>distress [0-42]; mean (SD)</td>
<td>2.3</td>
<td>(4.5)</td>
</tr>
<tr>
<td>fear-avoidance [0-42]; mean (SD)</td>
<td>14.1</td>
<td>(5.6)</td>
</tr>
<tr>
<td>kinesophobia [0-12]; mean (SD)</td>
<td>3.2</td>
<td>(3.5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Physical factors</strong></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>physical load at work [0-5]; mean (SD)</td>
<td>1.2</td>
<td>(1.5)</td>
</tr>
<tr>
<td>physical activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>less active than others</td>
<td>110</td>
<td>(39)</td>
</tr>
<tr>
<td>equally active</td>
<td>245</td>
<td>(42)</td>
</tr>
<tr>
<td>more active</td>
<td>226</td>
<td>(19)</td>
</tr>
<tr>
<td>inability to perform daily activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-30 days</td>
<td>184</td>
<td>(31)</td>
</tr>
<tr>
<td>1-12 months</td>
<td>61</td>
<td>(10)</td>
</tr>
<tr>
<td>sporting activities</td>
<td>230</td>
<td>(39)</td>
</tr>
<tr>
<td>cause of shoulder problem: sporting injury</td>
<td>29</td>
<td>(5)</td>
</tr>
</tbody>
</table>
Tables 2 and 3 show that for both measures of outcome, model composition was influenced by missing data (CCA vs. MI-5). When models were derived from imputed data, model composition diverged from the CCA model. For both measures of outcome predictors with lower predictive abilities in the CCA (i.e., rank order according to regression coefficient estimates) were not included in the MI-5 (e.g., concomitant lower extremity pain for shoulder disability and for pain intensity; sporting activities and higher physical workload). Predictors that were no part of the CCA model entered the MI-5 model for persistent shoulder disability (e.g., duration of complaints, somatisation, external locus of control and age) were included in the MI-5 model.
Tables 4, 5, 6 and 7 show the results of assessing model stability by the bootstrap model selection procedure. CCA and MI-5 models were not identified as the most frequently occurring combination of predictors for both outcome measures (Tables 4, 5, 6). Only the persistent shoulder disability MI-5 method was identical to its bootstrapped enhanced version (Table 7). Model selection frequencies for the most frequently selected models were uniformly low (ranging from 24.0% to 3.6%). Indicating on a large variability in model composition within the bootstrap replicate data sets. When fewer potential predictors are retained after the first step of the bootstrap model selection procedure, this variability seemed to decrease and model selection frequency increased.
Table 4. Complete case bootstrap model selection for the outcome persistent disability

<table>
<thead>
<tr>
<th>Predictors*</th>
<th>Most frequently selected models</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>inability to perform daily activities</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>both shoulders afflicted</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>shoulder complaints in the past year</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>concomitant lower extremity pain</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>both shoulders afflicted</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>more disability at baseline</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>concomitant lower back pain</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>older age</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>longer duration of complaints</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>acute onset</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Count 33 23 23 22 16

% 6.6 4.6 4.6 4.4 3.2

* only those predictors that appeared in ≥40% of the first bootstrap model selection step are presented

rank the order of appearance of predictors in the derived models arranged by their predictive ability (regression coefficient estimates)

MI·5+B the multiple imputation based bootstrap selected model (i.e., the most frequently occurring combination of predictors in 2500 replicate data sets of the second bootstrap model selection step)

MI·5 the multiple imputation based model using 5 imputed data sets was also the most frequently occurring combination of predictors in the 2500 bootstrap replicate data sets

Count the number of times the model was selected in the 2500 replicate data sets of the second bootstrap model selection step
Table 5. Imputed bootstrap model selection results for the outcome persistent disability

<table>
<thead>
<tr>
<th>Predictors*</th>
<th>Most frequently selected models</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 4 5 MI-5+B MI-5</td>
<td></td>
</tr>
<tr>
<td>concomitant lower back pain</td>
<td>X X X X X</td>
<td>1 1</td>
</tr>
<tr>
<td>longer duration of complaints</td>
<td>X X X X -</td>
<td>2 2</td>
</tr>
<tr>
<td>both shoulders afflicted</td>
<td>X X X X X</td>
<td>3 3</td>
</tr>
<tr>
<td>inability to perform daily activities</td>
<td>X X - X</td>
<td>4 4</td>
</tr>
<tr>
<td>higher scores for somatisation</td>
<td>X X X X X</td>
<td>5 5</td>
</tr>
<tr>
<td>higher scores for external locus of control</td>
<td>X X X X X</td>
<td>6 6</td>
</tr>
<tr>
<td>more disability at baseline</td>
<td>X X X X X</td>
<td>7 7</td>
</tr>
<tr>
<td>older age</td>
<td>X X - X X</td>
<td>8 8</td>
</tr>
<tr>
<td>shoulder complaints in the past year</td>
<td>X X - X</td>
<td>- -</td>
</tr>
<tr>
<td>concomitant lower extremity pain</td>
<td>- - - - -</td>
<td>- -</td>
</tr>
<tr>
<td>Count</td>
<td>91 77 56 54 52</td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>3.6 3.1 2.2 2.2 2.1</td>
<td></td>
</tr>
</tbody>
</table>

* only those predictors that appeared in ≥40% of the first bootstrap model selection step are presented

rank the order of appearance of predictors in the derived models arranged by their predictive ability (regression coefficient estimates)

MI-5+ B the multiple imputation based bootstrap selected model (i.e., the most frequently occurring combination of predictors in 2500 replicate data sets of the second bootstrap model selection step)

MI-5 the multiple imputation based model using 5 imputed data sets was also the most frequently occurring combination of predictors in the 2500 bootstrap replicate data sets

Count the number of times the model was selected in the 2500 replicate data sets of the second bootstrap model selection step
Table 6. Complete case bootstrap model selection for the outcome persistent pain

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Most frequently selected models</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>longer duration of complaints</td>
<td>X X X X X</td>
<td>1 3</td>
</tr>
<tr>
<td>concomitant lower back pain</td>
<td>X X X X X</td>
<td>2 2</td>
</tr>
<tr>
<td>both shoulders afflicted</td>
<td>X X - X X</td>
<td>3 4</td>
</tr>
<tr>
<td>concomitant upper extremity pain</td>
<td>X X X -</td>
<td>4 6</td>
</tr>
<tr>
<td>shoulder complaints in the past year</td>
<td>X - X X</td>
<td>5 -</td>
</tr>
<tr>
<td>sporting injury*</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>inability to perform daily activities*</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>sporting activities*</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>higher physical workload*</td>
<td></td>
<td>8</td>
</tr>
</tbody>
</table>

Count 120 96 58 47 37

% 24.0 19.2 11.6 9.4 7.1

* predictors that appeared in <40% of the first bootstrap model selection step are not used in the second model selection step

rank the order of appearance of predictors in the derived models arranged by their predictive abilit (regression coefficient estimates)

MI-5+B the multiple imputation based bootstrap selected model (i.e., the most frequently occurring combination of predictors in 2500 replicate data sets of the second bootstrap model selection step)

MI-5 the multiple imputation based model using 5 imputed data sets was also the most frequently occurring combination of predictors in the 2500 bootstrap replicate data sets

Count the number of times the model was selected in the 2500 replicate data sets of the second bootstrap model selection step
Table 7. Imputed bootstrap model selection results for the outcome persistent pain

<table>
<thead>
<tr>
<th>Predictors*</th>
<th>Most frequently selected models</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>sporting injury</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>longer duration of complaints</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>concomitant lower back pain</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>both shoulders afflicted</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>inability to perform daily activities</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>higher level of education</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>shoulder complaints in the past year</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>concomitant upper extremity pain</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>higher physical workload</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Count</td>
<td>163</td>
<td>158</td>
</tr>
<tr>
<td>%</td>
<td>6.5</td>
<td>6.3</td>
</tr>
</tbody>
</table>

* predictors that appeared in <40% of the first bootstrap model selection step are not used in the second model selection step

rank the order of appearance of predictors in the derived models arranged by their predictive ability (regression coefficient estimates)

MI-5+B the multiple imputation based bootstrap selected model (i.e., the most frequently occurring combination of predictors in 2500 replicate data sets of the second bootstrap model selection step)

MI-5 the multiple imputation based model using 5 imputed data sets was also the most frequently occurring combination of predictors in the 2500 bootstrap replicate data sets

Count the number of times the model was selected in the 2500 replicate data sets of the second bootstrap model selection step
Model performance

Table 8 presents the performance of the models derived with the four methods for both outcome measures. The slopes of the calibration plots ranged from 0.973 to 1.077, which indicates good calibration. Explained variance ranged from 8.8% to 12.0% for disability and from 13.5% to 18.8% for pain. The apparent c-indices varied between 0.645 and 0.667 for disability and between 0.684 and 0.717 for pain intensity. CCA models were more optimistic compared to the other models. Following adjustment for overoptimism the corrected c-indices were within the range of 0.639–0.646 for persistent shoulder disability and within the range of 0.667–0.688 for persistent shoulder pain.

DISCUSSION

Prognostic research aims at identifying the ingredients that contribute to a correct prognosis for a specific subgroup of patients. Though, finding a stable set of predictors that can consistently be used in a broad patient population proves to be difficult. Several methodological issues (missing data and model stability) which are not accounted for by the standard statistical methodology are expected to complicate this matter. We showed that accounting for missing data by MI and providing information on model stability by bootstrapping are instructive methods when deriving a prognostic model.

In the standard statistical methodology the use of a backward or forward selection strategy has been criticized. It may result in overoptimistically estimated regression coefficients, omission of important predictors and random selection of less important predictors. Derived models may therefore be unstable. Research has
Table 8. Measures of model performance

<table>
<thead>
<tr>
<th></th>
<th>Persistent shoulder disability</th>
<th>Persistent shoulder pain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCA</td>
<td>MI-5</td>
</tr>
<tr>
<td>calibration slope</td>
<td>0.978</td>
<td>0.978</td>
</tr>
<tr>
<td>R²ₐ</td>
<td>0.119</td>
<td>0.120</td>
</tr>
<tr>
<td>A_c</td>
<td>0.666</td>
<td>0.667</td>
</tr>
<tr>
<td>95% CI</td>
<td>.616–.715</td>
<td>.627–.710</td>
</tr>
<tr>
<td>optimism</td>
<td>0.027</td>
<td>0.022</td>
</tr>
<tr>
<td>O_C</td>
<td>0.639</td>
<td>0.646</td>
</tr>
</tbody>
</table>

A_c                apparent c-index  
B                  bootstrapping base don a complete case data set  
CCA               complete case analysis  
MI-5%+B           multiple imputation combined with bootstrapping  
MI-5              multiple imputation using 5 imputation files  
O_C               optimism corrected c-index  
Optimism          estimation of the overoptimism  
R²ₐ               explained variance (Nagelkerke’s R-squared)  
95% CI            95% confidence interval
focused on how to derive stable models. One frequently used method is the bootstrapping approach suggested by Austin and Tu.\textsuperscript{35} It considers the strength of evidence that identified variables are truly important predictors in resampled data. Although this approach is often claimed to reduce model instability,\textsuperscript{8,10,14,35,36} separating strong from weak predictors was shown to perform comparative to automated backward elimination in identifying the true regression model.\textsuperscript{37} Furthermore, this approach has limited abilities when there is a high number of potential prognostic factors. For these situations a modified bootstrapping procedure was suggested.\textsuperscript{11} Our study showed that the application of this two-step bootstrap model selection procedure provides valuable information on model stability.

As frequently described, model size and model composition are also affected by missing data. Especially in standard statistical methodology where subjects with missing values on any of the recorded variables are omitted from analysis. When missing data does not depend on observed or unobserved measurements (Missing Completely At Random, MCAR), this leads to loss of costly gathered information, decreased statistical power, altered associations between predictors and therefore differences in model composition.\textsuperscript{12,13,38–41} In this context our study findings formed no exception. Model composition varied as a result of whether cases with missing data were omitted from analyses (CCA) or whether the values of the missings were estimated using MI. Since missing values appeared to be related to other observed information, the MCAR condition did not hold and CCA was expected to be biased. Most of the missing data was observed in the outcome because participants did not consent to follow-up. As subjects lost to follow-up showed more often an acute onset (47% versus 37%), were younger (mean difference of 7 years) and the variable age is
included in the MI model for the outcome measure persistent shoulder disability, it is plausible to assume that these missings are MAR. For that reason, accounting for missing data by MI using 5 imputed data sets was in our multivariate data setting the most optimal choice to reduce the uncertainty in model derivation caused by missing values. The use of even more data sets in the imputation routine is possible (up to 20), however 5 was shown to be an sufficient number in order to get stable results.\textsuperscript{30} Yet the addition of a bootstrap model selection procedure showed that the MI-5 model might still be unstable. A possible source for this instability might be the suboptimal variable selection procedure applied in the MI-5 procedure. However, how to optimally perform variable selection in multiple imputed data is still a subject of discussion.\textsuperscript{42} As illustrated by our study, the bootstrap model selection procedure may provide valuable additional information on model stability when deriving a prognostic model in multiple imputed data. To study the effects of accounting for missing data and incorporating model stability we used a large clinical data set in which we empirically evaluated methods of deriving a prognostic model. By this, the uncertainties researchers commonly face when knowledge of the true predictors of outcome is lacking, were illustrated. Furthermore, the practical utility of the additional information provided by the bootstrap model selection procedure in prognostic modelling is demonstrated. Though results need to be interpreted with caution, as our approach limits us from identifying a superior methodology. Although performance parameters for each derived model are presented, these play no role in the decision on the superiority of a certain method. They only show that the performance of all derived models was comparable to that from existing clinical prediction rules on shoulder pain.\textsuperscript{6, 15} For deciding on the superiority of a certain method, a simulation study in which true predictors and noise variables are assigned would be needed. Such data is not presented by this study.
Conclusions

Our study showed that in this particular dataset of shoulder pain patients, model composition varied as a result of how missing data was handled. Furthermore, the bootstrap model selection routine gave additional information on model stability.
References


29. Harrell FE Jr: Checklist for Authors: Statistical Problems to Document and to Avoid. Vanderbilt University Department of biostatistics 20 Dec 2007. [http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/ManuscriptChecklis


LOGISTIC REGRESSION AND CART ANALYSIS

development and validation of a clinical prediction model for persistent pain by logistic regression and CART analysis

Vergouw D, Heymans MW, Kuijpers T, van der Windt DAWM, van der Horst HE, de Vet HCW. Development and validation of a clinical prediction model for persistent shoulder pain by logistic regression and CART analysis. Submitted for publication
ABSTRACT

Objectives: Prognostic models are usually derived by regression analyses, but recursive partitioning can be regarded as a good alternative method. Previous comparisons of both techniques have resulted in inconclusive conclusions about predictive ability since performance and internal validation are hardly ever properly compared.

Methods: To study which method was better suited for the derivation of a prediction model predicting persistent shoulder pain three months after initial consultation in primary care, recursive partitioning by the Classification And Regression Tree (CART) algorithm was compared with a logistic regression model regarding model composition, model performance (Area Under the Curve (AUC), explained variance $R^2_N$) and internal validity (bootstrapping). Models were derived in a cohort of 587 patients consulting with a shoulder problem in general practice in the Netherlands. Potential predictors included socio-demographic variables, characteristics of the pain problem, physical activity and psychosocial factors.

Results: Results showed that model composition differed, although model performance in the original dataset was comparable (equal $R^2_N$ of 19%, an AUC of 0.72 for logistic regression vs. 0.70 for CART). Applied to comparable subjects, the CART model appeared more overoptimistic than the regression model.

Conclusion: We conclude that the logistic regression model is better suited for the derivation of prediction models.
INTRODUCTION

The derivation of prognostic models and the use of prediction rules have become increasingly popular in the last decade. These models and rules provide a quantitative estimation of the risk of a particular outcome for individual patients. Clinicians can use these estimates when making decisions regarding treatment.¹

Usually prognostic models are derived by applying regression analysis, however alternative methods do exist. One of such methods is recursive partitioning or decision tree analysis.² With this technique a tree-like diagram is created which can be used to categorise individual patients according to their expected event outcome by consecutive simple yes/no questions. Regression models often require the conversion of the final model into a clinical easy to use prediction tool such as a score chart. In contrast to this, recursive partitioning is said to be more clinically intuitive since resulting models closely resemble clinical reasoning and decision making without further adaptation. Furthermore, recursive partitioning provides the methodological surplus of automatic interaction effect identification, where in regression analysis the identification of interaction effects is more complex and requires researcher and clinician expertise.³

From the above, recursive partitioning can be regarded as a promising alternative technique for the derivation of a clinical prediction rule. Several studies have compared recursive partitioning with regression models. However, results are inconclusive because of methodological heterogeneity.⁴ Mostly comparisons are improper or incomplete because 1) only variables identified as predictors of outcome (i.e. model composition) were compared which provided no information on the predictive ability of
the models derived by both methods, 2) model performance (defined by sensitivity & specificity or misclassification error) was valued against an arbitrarily chosen cut-off value and therefore results were largely data specific, or 3) the predictive abilities (as indicated by the Area Under the receiver operating Curve, AUC) are only estimated in the original data while predictive abilities typically deteriorate when the model is applied in new subjects due to optimistic model estimates. A complete comparison of both methods should therefore not only include performance measures estimated in the original data, but also measures estimated in a validation sample. By dividing the original dataset into a large derivation set and a smaller validation set, some studies have tried to validate their model. However, this one split-sample approach is inefficient for model building, since only a part of the study sample characteristics are represented in the model derivation dataset. The remaining and thus not the complete dataset is used to test and validate the model. This can seriously affect the generalizability of the derived model in new patients. Therefore, a form of repeated split-sampling (i.e., cross-validation or bootstrap resampling) is recommended to first internally validate the model. These methods use the complete original dataset to derive and validate the model, which may increase their generalizability. Only a few studies have incorporated such an internal validation procedure when comparing prediction models derived by recursive partitioning and regression models. However, results remain inconclusive as one study concluded that regression models had superior performance compared to recursive partitioning while the two others concluded that both methods had comparable performance. Heterogeneity in the number of cross-validation repetitions (i.e. Austin used 1000 repetitions, Gansky used 5 repetitions, and Rosenfeld and Lewis repeatedly split the data randomly in a 90% derivation and 10% validation set until all cases had been used in the derivation set)
could very well account for this inconsistency since Efron has shown accuracy estimates to be highly unstable over cross-validation repetitions.\textsuperscript{14}

In the current study we compare recursive partitioning with a logistic regression model and moreover, we compared both methods at all steps of model development. Next to model composition and model performance we assess the internal validity for models predicting persisting shoulder pain 3 months after first primary care consultation.

\section*{METHODS}

\subsection*{Data}

Data from the Dutch Shoulder Study (DSS) cohort was used.\textsuperscript{15} Patients (n=587) who consulted their general practitioner (GP) with a new episode of shoulder disorders were enrolled in this cohort. A new episode was defined as: no GP consultation or treatment received for the affected shoulder in the preceding three months. Exclusion criteria were: severe physical conditions (i.e., fractures or dislocation in the shoulder region, rheumatic diseases, neoplasms, neurological of vascular disorders) or cognitive limitations which may hamper the completion of questionnaires. The ethics review board of the VU University medical centre approved the study protocol.

\subsection*{Outcome measures}

We focused on the outcome measure persistent shoulder pain intensity (Numeric Rating Scale; 0-10). To define ‘persistent pain intensity’ baseline pain scores were subtracted from follow-up pain
scores. Patients who improved less than 50% were denoted as having persistent pain.\textsuperscript{16} Outcomes were measured three months after enrolment.

**Candidate prognostic factors**

Based on a systematic review of the literature\textsuperscript{17} candidate predictors were selected, including demographic variables, characteristics of the shoulder pain problem, physical and psychological factors. The following questionnaires were used to gather information on psychological factors: the Pain Coping and Cognition List (PCCL\textsuperscript{18}) to measure pain coping, and catastrophizing, the 4 Dimensional Symptom Questionnaire (4DSQ\textsuperscript{19}) to assess anxiety, depression, distress, somatisation, the Fear-Avoidance Beliefs Questionnaire (FABQ\textsuperscript{20}) to measure fear-avoidance and the Tampa Scale for Kinesiophobia (TSK\textsuperscript{21,22}) to assess kinesophobia. Within 10 days after consulting the GP, participants completed a baseline questionnaire to assess potential predictors.

**Regression analysis**

The linearity assumption was checked for all continuous candidate prognostic factors. When the relationship between variables and outcome did not resemble linearity, variables were categorized (3 categories) or dichotomized. Variables were checked for (multi)collinearity using Pearson’s r, given that correlated variables can disturb variable selection in multivariable regression\textsuperscript{23}. In case two variables showed a correlation (Pearson’s R) of 0.5 or more, the variable which could be most easily obtained in clinical practice by the physician was retained. Subsequently a multivariable logistic regression analysis with a backward selection strategy and a stopping rule of $\alpha=0.157$ was performed. This significance level is known as the Akaike Information Criterion (AIC).\textsuperscript{24,25} No interaction
terms were examined in this analysis.

**Recursive partitioning**

The binary recursive partitioning of the training data is performed by the Classification And Regression Trees (CART) algorithm developed by Breiman and colleagues.\textsuperscript{26} This technique requires no a priori data preparation like data normalisation, and checking for linearity assumptions. Furthermore, it can handle numerical as well as categorical data and interaction terms are identified automatically. By the application of the CART method, the original data is successively divided into subgroups. For each split all possible splitting variables and all possible cut-off values are examined. The one split that results in two subgroups that are most different with respect to the outcome according to a predefined splitting criterion (Gini-index, for more details see Breiman et al.\textsuperscript{26}), is chosen as the best split and will be used in the decision tree. By this a parent node will be divided into two child nodes. This procedure is continued until no further splits can be made and thus a maximum sized decision tree is created. Since maximum sized trees are mostly overfitted models, they need to be reduced in size by snipping of branches, which is called pruning. Pruning is a method to adjust for overoptimism.\textsuperscript{26,27} The standard pruning method was applied in which an optimally sized tree can be found by studying the relationship between tree-size and tree accuracy (i.e., misclassification). For a more detailed description of pruning see Breiman et al.\textsuperscript{26} or Therneau and Atkinson\textsuperscript{27}.

**Comparison of methods**

To ensure a fair comparison between methods, the same set of candidate predictors was used in both CART and regression analysis. Differences between both methods were assessed by
1) studying the agreement on which factors were determined as prognostically important, 2) by comparing model performance measures such as 2a) how well the models distinguish between patients with and without persistent symptoms, i.e., discrimination as indicated by the c-index which for binary outcomes is identical to the area under the ROC curve and therefore in the latter will be called the AUC, 2b) the agreement between predicted and observed probabilities for each model i.e., calibration indicated by the slope of the calibration plot, and 2c) the explained variance of both models as indicated by Nagelkerke’s R²N, and 3) an estimation of the models’ overoptimism or how well they will perform when applied in new subjects, also called internal validation.

**Internal validation**

Regression and CART models typically overfit the original data. This means that in general their performance is better in the data set in which they are developed compared to their performance in another similar dataset. To estimate this overfit or overoptimism, a technique commonly used for internally validating regression models, bootstrapping¹⁴ was employed using 200 bootstrap replicate datasets. In its general form this technique uses the AUC as a measure of model performance to be corrected for overoptimism and contains the following steps:¹⁴

1) Build the prognostic model in the original dataset, determine its predicted probabilities and estimate its AUC\(_\text{original}\);  
2) Draw a random bootstrap sample with replacement from the original dataset. This bootstrap sample has the same size as the original dataset;  
3) Re-estimate in the bootstrap sample the regression coefficients of the original model, determine its predicted probabilities and
estimate the AUC\textsubscript{bootstrap};

4) Freeze the in the previous step re-estimated regression coefficients, apply them in the original dataset and estimate the AUC\textsubscript{test} (the original data has become a test set now);

5) Calculate the overoptimism by subtracting

\[\text{AUC}_{\text{bootstrap}} - \text{AUC}_{\text{test}} = \text{AUC}_{\text{overoptimism}};\]

6) Repeat steps 1-5 about 200-250 times and average the overoptimism;

7) Correct the AUC\textsubscript{original} for the average overoptimism by substracting AUC\textsubscript{original} – average overoptimism.

Because of the differences between the CART and regression methodology, the previous steps 3 and 4 need some further elucidation when overoptimism is determined for the CART model. Since CART analysis does not generate regression coefficients, the AUC\textsubscript{bootstrap} is determined by applying the original CART model in the bootstrap dataset and obtaining the predicted probabilities from the node proportions in each endnode. Subsequently, this bootstrap CART model is frozen in step 4 and applied to the original data to obtain the AUC\textsubscript{test}. These measures are subsequently used in the next steps to estimate the overoptimism of CART models.

**Software**

All analyses were performed using the R-statistics software (version 2.4.0). The R Design package was used for the logistic regression, the rpart package was used for the CART analysis and additional routines were developed for applying the bootstrap internal validation routine.
Figure 1. The CART decision tree for predicting persistent shoulder pain intensity three months after first consultation in primary care. Each node contains the number of participants in that node and the node proportions. In each endnode (squares) the probability of having persistent symptoms is depicted.
Table 1. Model composition, regression coefficients and odds ratios for the regression model predicting persistent shoulder pain three months after first consultation in primary care.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>category</th>
<th>β</th>
<th>(SE)</th>
<th>OR</th>
<th>(95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>persistent shoulder pain*</td>
<td>persistent/recovered</td>
<td>-1.313</td>
<td>(0.523)</td>
<td>0.30</td>
<td>(0.11–0.83)</td>
</tr>
<tr>
<td>sporting injury</td>
<td>yes</td>
<td>-0.962</td>
<td>(0.234)</td>
<td>2.62</td>
<td>(1.65–4.14)</td>
</tr>
<tr>
<td>concomitant lower back pain</td>
<td>yes</td>
<td>0.54</td>
<td>(0.223)</td>
<td>1.71</td>
<td>(1.03–2.85)</td>
</tr>
<tr>
<td>longer duration of complaints</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6–11 weeks</td>
<td></td>
<td>0.537</td>
<td>(0.260)</td>
<td>1.71</td>
<td>(1.03–2.85)</td>
</tr>
<tr>
<td>&gt;11 weeks</td>
<td></td>
<td>0.857</td>
<td>(0.238)</td>
<td>2.36</td>
<td>(1.48–3.76)</td>
</tr>
<tr>
<td>both shoulders afflicted</td>
<td>yes</td>
<td>0.666</td>
<td>(0.301)</td>
<td>1.95</td>
<td>(1.08–3.51)</td>
</tr>
<tr>
<td>inability to perform daily activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–30 days</td>
<td></td>
<td>-0.617</td>
<td>(0.223)</td>
<td>0.54</td>
<td>(0.35–0.83)</td>
</tr>
<tr>
<td>1–12 months</td>
<td></td>
<td>-0.328</td>
<td>(0.337)</td>
<td>0.72</td>
<td>(0.37–1.39)</td>
</tr>
<tr>
<td>concomitant upper extremity pain</td>
<td>yes</td>
<td>0.347</td>
<td>(0.214)</td>
<td>1.41</td>
<td>(0.93–2.15)</td>
</tr>
<tr>
<td>sporting activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–3 hours/week</td>
<td></td>
<td>-0.155</td>
<td>(0.226)</td>
<td>0.86</td>
<td>(0.55–1.33)</td>
</tr>
<tr>
<td>&gt;3 hours/week</td>
<td></td>
<td>0.160</td>
<td>(0.313)</td>
<td>1.17</td>
<td>(0.63–2.17)</td>
</tr>
</tbody>
</table>

* = outcome measure
β = indicates regression coefficient estimates
SE = indicates standard error
OR = indicates odds ratio
RESULTS

Model composition

The composition for the regression and CART models are respectively shown in Figure 1 and Table 1. The total number of included predictors was equal for both methods (both regression analysis and CART included seven predictors). From these, both models share three predictors (i.e., duration of complaints, concomitant low back pain and both shoulder afflicted) which are independently identified as important predictors of outcome in each individual model (indicated by larger regression coefficients or splits placed higher in the tree hierarchy). However, the predictor identified by the regression analysis as the most strongly associated predictor with outcome (i.e., having a sports injury) is not included in the CART model.

Figure 2. Calibration plot showing the observed versus the predicted probabilities for persistent shoulder pain intensity three months after the first consultation in primary care for the regression model (black squares) and the CART decision tree (grey triangles).
Model performance and internal validation

The calibration plot for both models by using the original data is provided in Figure 2. The range of predicted probabilities was 0.2 to 0.8 for the logistic regression model. The range for the CART model was somewhat smaller, i.e., 0.3 to 0.8. Furthermore, this model is not able to predict probabilities within a range of 0.4 to 0.6. Thus the CART model is adept at identifying clinical subgroups of subjects at very low or high risk of adverse shoulder pain intensity where the logistic regression model shows a larger variation in identifying individual patient risk. No deviations from the ideal calibration (i.e., the 45° line) were observed.

Other model performance measures are depicted in Table 2. The explained variance of both methods was equal with a value of 19%. Discriminative abilities in the original dataset were better for the regression model (AUC=0.72) compared to the CART model (AUC=0.70). After correction for overoptimism as estimated by the internal validation procedure, a larger difference in AUC between

Table 2. Model performance measures for the logistic regression model (LR) and the CART decision tree predicting persistent shoulder pain intensity three months after first consultation in primary care in the original dataset and corrected for overoptimism.

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>CART</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2_N$</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>c-index (AUC)</td>
<td>0.72</td>
<td>0.70</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.67-0.77</td>
<td>0.66-0.75</td>
</tr>
<tr>
<td>Optimism</td>
<td>0.04</td>
<td>0.17</td>
</tr>
<tr>
<td>AUC Optimism corrected</td>
<td>0.68</td>
<td>0.53</td>
</tr>
</tbody>
</table>

$R^2_N$ indicates explained variance
95% CI indicates 95% confidence interval
c-index indicates model discrimination (AUC)
both methods occurred. This can be explained by the larger overoptimism for the CART model compared to the logistic regression model (overoptimism for CART 0.17 vs. 0.04 for logistic regression).

**DISCUSSION**

This paper studies the performance of prognostic models developed with logistic regression and CART analysis. Both methods were compared at several important steps of model development, i.e., model composition, model performance, and internal validation. Because of large methodological heterogeneity, previous comparisons between recursive partitioning and regression models provided inconclusive results. According to proponents of recursive partitioning, the distinctive tree-like structure of the resulting models makes them clinically more useful than regression analysis based models. However, next to being clinical useful, prognostic models foremost need to be valid and accurately predict outcome. By comparing decision trees derived by the CART methodology with prognostic models derived by logistic regression analysis, our study showed that both models performed equally well in the original data set, but that the regression based model was more (internally) valid than the CART model.

**Model composition**

Model composition differed between the CART and regression models. One explanation for the observed differences in model composition could be the difference in predictor selection strategy of both methods. In CART a number of algorithms (i.e., chi-square, misclassification error, Gini index, cross-entropy) can be chosen in order to grow a tree. In our study, the frequently employed and by
Breiman et al.\textsuperscript{26} preferred Gini-index was used to select predictors in CART. In contrast, predictor selection in our logistic regression analysis was based on p-values. This means that where CART identified unequally distributed outcome subgroups by automatically establishing optimal cut-off points, logistic regression analysis tried to select a subset of relevant predictors by studying whether adding or removing predictors resulted in a predetermined effect size. These different predictor selection criteria might explain the differently composed models. Furthermore, the observed differences in model composition could be explained by the fact that predictor selection strategies are generally known to give inconsistent results independent of the strengths of associations between predictors and outcome.\textsuperscript{28}

**Model Performance**

Model performance measures (i.e., discrimination as indicated by the AUC and explained variance as indicated by Nagelkerkes’ $R^2$) obtained in the original dataset showed strong resemblance for recursive partitioning and regression models. These findings are in agreement with other studies which compared both methods and found comparable performance.\textsuperscript{6,9} However, based on results obtained in the original dataset, it cannot be concluded that both CART and regression models will perform equally well when applied in new subjects. Model performance measures are known to typically deteriorate when applied in new subjects, since CART and regression models usually overfit the original dataset. Therefore, a validation routine has to be employed to correct for this overfit and derive a more generalizable estimation of the models’ performance. Whilst such a routine (i.e., cross-validation or bootstrap resampling) has become standard in regression based prognostic research, in recursive partitioning it is rarely employed properly. We only came
across one study\textsuperscript{4} that used a validation routine up to current statistical standards. Our bootstrap internal validation routine was comparable to this method and the results of our study seem to strengthen the findings reported by Austin, i.e., CART models do overfit the original dataset more than regression models do. Thus, even though the CART model has already been adjusted for overoptimism by the pruning routine, it showed a greater reduction of the AUC when applied to comparable subjects, than the uncorrected regression model.

**Strengths and weaknesses**

The current study compared model composition, model performance and model validity for both a CART and a regression model. This can be regarded as one of the strengths of this study, since previous studies often only studied these aspects separately. The addition of the internal validation routine in particular showed the advantage of logistic regression over CART regarding generalizability. Although validation in new subjects is preferable over internally validating a model by using the derivation dataset, our results on generalizability are strengthened by a recent study in which logistic regression models were shown to be more accurate than CART models when applied in new subjects.\textsuperscript{31} Furthermore, this study was performed using a large empirical dataset which may add to the generalizability of our derived models. One possible weakness of our study is the handling of missing data. It is known that excluding cases with missing values from analysis might result in biased prognostic models.\textsuperscript{32} Therefore, the use of a complete case analysis as employed in our regression analysis, is strongly discouraged. CART however, can only use observed data when selecting each split and thus for each split CART performs also a complete case analysis.
Although the results of both methods in our study might be biased because of missing values, we are confident that missing data did not affect the comparison procedure of both methods in this study.

**Conclusions**

Initial model appeared to be equal for both CART and regression models when studied in a large empirical dataset. However, a bootstrapping internal validation routine showed that CART models did overfit the original dataset more than regression models did. Therefore, we conclude that the CART model does not perform as well as the logistic regression models did. Therefore, we conclude that the CART model does not perform as well as the logistic regression model for predicting persistent shoulder pain, especially with regard to the internal validity of the model. Our study adds to the evidence that logistic regression is better suited for the derivation of prognostic models than CART.
References


PREDICTION IN GENERAL PRACTICE

comparing clinical consensus from a Delphi procedure with a statistical scoring system

Published as:

ABSTRACT

Objectives: In prognostic research, prediction rules are generally statistically derived. However the composition and performance of these statistical models may strongly depend on the characteristics of the derivation sample. The purpose of this study was to establish consensus among clinicians and experts on key predictors for persistent shoulder pain three months after initial consultation in primary care and assess the predictive performance of a model based on clinical expertise compared to a statistically derived model.

Methods: A Delphi poll involving 3 rounds of data collection was used to reach consensus among health care professionals involved in the assessment and management of shoulder pain.

Results: Predictors selected by the expert panel were: 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. When tested in a sample of 587 primary care patients consulting with shoulder pain the predictive performance of the two prognostic models based on clinical expertise were lower compared to that of a statistically derived model (Area Under the Curve, AUC, expert-based dichotomous predictors 0.656, expert-based continuous predictors 0.679 vs. 0.702 statistical model).

Conclusions: The three models were different in terms of composition, but all confirmed the prognostic importance of symptom duration, baseline level of shoulder disability and multisite pain. 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.
INTRODUCTION

A clinical prediction rule is a simple tool which uses a combination of early signs or symptoms to provide a quantitative estimate of the absolute risk of particular outcomes for individual patients. Often the outcome is the individuals’ expected course of an illness (prognosis), however clinical prediction rules can also be developed for predicting presence of a disease (diagnosis) or for predicting an individuals’ response to a particular intervention. The obtained estimations may subsequently be used by clinicians for the provision of patient information or to support decisions regarding treatment and referral.

Before a prediction rule can be implemented in clinical practice it ideally needs to be developed, validated and analysed for impact. In prognostic research, prediction rules are generally derived by logistic or Cox regression models. With these statistical models, predictors are selected from a larger pool of potential predictors which is frequently established prior to model derivation and originates from previous literature or expert recommendations. Selection is frequently based on forward or backward regression analysis in combination with a predefined p-value. However, it is not uncommon that prediction models derived by these methods perform poorly, with composition and predictive performance strongly dependent on characteristics of the derivation dataset. Especially for the prediction of non-specific musculoskeletal symptoms, the identification of good prediction models has appeared to be difficult.

In order not to miss potential predictors, prognostic researchers tend to gather an excessive amount of data, after which a smaller set of predictors is selected using statistical methods. Many prognostic models, however, especially in the area of musculoskeletal conditions consists of studies incorporating small sample sizes that
are not in agreement with the suggested potential predictor to subject ratio\textsuperscript{4} required for subsequent statistical analyses. Under these conditions, predictor selection by using statistical methods is known to yield unstable results independent of the strength of the association between predictor and outcome.\textsuperscript{2} This may hamper the derivation of clinical useful prediction models with good practical performance and as a result has potential to be associated with invalid results in subsequent analysis (e.g., model validation).

When statistically deriving a prediction model, the contribution of clinical expertise prior to model derivation is often minimal. As a result, potential predictors obvious to clinicians might be overlooked. Therefore, the incorporation of clinical knowledge in the early phase of predictor selection can be of great importance. A technique known as a Delphi procedure,\textsuperscript{5} is believed to be an effective and reliable way of obtaining expert-based knowledge\textsuperscript{6,7} and can be applied in prognostic research.\textsuperscript{8,9} In this procedure, a group of experts responds anonymously to a series of subsequent questionnaires. Results are fed back to the panel in order to reach consensus. A potential advantage of this is that through the anonymous nature of the Delphi, negative group interactions (such as dominant group members forcing their beliefs onto the entire group) are eliminated.\textsuperscript{5}

In the present study we aimed to reach consensus among clinicians and experts regarding predictors that are most important for predicting persistent shoulder pain three months after initial consultation in primary care. A Delphi procedure with an expert panel of health care professionals involved in the assessment and management of shoulder pain was used to identify these key predictors. In a first step to determine the quality of these expert-based selected predictors two clinical/expert-based prognostic models were constructed and their predictive performance was compared
with a third statistically derived prognostic model. Differences in model composition between the statistical and clinical models were expected and since the statistical model was modelled for the chosen data set, predictive performance was expected to be better for the statistical model. These results will allow us to comment on whether the clinical models are an appropriate comparator model for future studies investigating predictive performance in a new sample (e.g., where statistical models often falter).

METHODS

Delphi procedure

*Expert panel selection*

A multidisciplinary panel with members involved in or having thorough knowledge of shoulder pain in clinical practice was formed. We invited general practitioners, physiotherapists, orthopaedists and manual therapists from the United Kingdom as well as from the Netherlands to participate in the Delphi study. Expertise was defined by a) membership of a professional organisation combined with specific expertise in shoulder conditions (e.g., members of the U.K. Primary Care Rheumatology Society or the Dutch College of General Practitioners), b) being involved in guideline development or clinical research on shoulder pain or c) having a special interest and significant experience in treating shoulder conditions.

In order to obtain reliable results, a Delphi panel minimally needs to consist of 10 to 15 experts. More participants will add to the reliability, but will complicate the procedure. We aimed to compose an expert panel of 20 members, a number which is commonly seen in consensus based research. Accounting for non-response, we approached 52 experts in the area of shoulder
symptoms. All were provided with an information letter explaining the aims, procedures and requirements of the Delphi study.

First round: ranking potential predictors
Similar to previous prognostic consensus studies, first a list of potential predictors based on a systematic review was composed. In the first round, all panel members were presented with this list, which was sub-grouped in 7 categories (demographic, general health, characteristics of the shoulder symptom, pain related, psychological, social and physical load and activity). The panel members were asked to score the importance of each potential predictor on a 5 point Likert scale (i.e., 1= not at all predictive, to 5= highly predictive). When panel members felt that important predictors were missing from the provided list, they were encouraged to suggest additional potential predictors. Based on a summation of these scores all potential predictors were ranked according to their predictive ability. Newly suggested predictors were added to the list and arranged by the frequency with which they were suggested.

Second round: re-evaluation of predictive abilities
The panel received feedback on the results of round one, and was subsequently asked in round two to rank the 10 most important potential predictors. This was done by rewarding the strongest predictor with 10 points and the weakest with 1 point. Hence all potential predictors were re-evaluated and arranged according to their predictive performance (the total of points rewarded to each potential predictor).

Third round: consensus on the 10 most important predictors
In this third round panel members were asked whether or not they agreed on the 10 most important predictors from the second round.
In case of disagreement, panel members were able to alter the selection by replacing a maximum of 3 predictors. Predictors could be eliminated from the selection or be replaced by others. In order to reduce the replacement options and come to consensus more easily, predictors could only be replaced by the 20 most predictive factors from round two. When eliminating or replacing any predictor, panel members were asked to provide a rationale for their decision. To reach consensus we a priori determined that at least 90% of the panel had to agree on all the predictors selected. If predictors were replaced or changed, participant consensus on the updated predictors was re-evaluated as part of round three. Based on argumented alterations a new selection was formed (i.e., predictors with <90% inclusion agreement were replaced by the most frequently mentioned replacement options) and together with these arguments were presented again for consensus.

**Predictive performance of the expert-based model**

*Data*

The prediction models that were comprised of the predictors selected by our expert panel were evaluated and compared to a previously derived statistical prognostic model\(^{13}\) using an existing data set. Data used for this purpose came from the Dutch Shoulder Study,\(^ {14} \) a cohort of 587 subjects consulting with a new episode of shoulder pain (i.e., had not consulted the GP or received treatment for the current shoulder problem in the previous three months) in general practice. Exclusion criteria of this cohort were: severe physical or psychological conditions (i.e., fractures or luxation in the shoulder region, rheumatic disease, neoplasms, neurological or vascular disorders, dementia) or cognitive limitations which would hinder the completion of a written questionnaire. The study was approved by the Medical Ethics Committee of the VU University Medical Center,
Amsterdam, The Netherlands.

**Outcome**

Persistent shoulder pain was defined by subtracting baseline scores (numeric scale 0-10) from follow-up scores. Subjects who improved less than 50% were indicated as having persistent shoulder pain. This definition of persistence was previously shown to be the minimal important change and was therefore used as cut-off value.\(^{13}\) Outcome was measured at three months after initial consultation by a postal questionnaire.

**Variables in the dataset**

Collected data included demographic, shoulder pain related, physical and psychosocial factors, which were on average recorded 10 days after initial primary care consultation. Next to single questions, validated questionnaires were used to gather data.\(^{14}\) Questionnaires used were: the Shoulder Disability Questionnaire (SDQ\(^{15}\)) to assess the shoulder symptom severity (potential range: 0-100 points), the Pain Coping and Cognition List (PCCL\(^{16}\)) to measure coping with pain (1-6 points), catastrophizing (1-6 points), internal (1-6 points) and external locus of control (1-6 points), the Four-Dimensional Symptom Questionnaire (4DSQ\(^{17}\)) to assess anxiety (0-24 points), depression (0-12 points), distress (0-32 points) and somatisation (0-32 points), the physical activity scale of the Fear-Avoidance Beliefs Questionnaire (FABQ-PA\(^{18}\)) to measure physical activity related fear-avoidance (0-42 points) and the Tampa Scale for Kinesiophobia (TSK\(^{19,20}\)) items no. 1 and 9 to assess kinesiophobia (0-12 points).

**Analysis**

In order to retrieve information of all expert-based selected
predictors, individual questionnaire items had to be used or combined. Consequently, with this information, we explored two different possibilities to derive an clinical/expert-based prognostic model. These methods will be compared and are explained below. Both clinical models consisted of the ten Delphi selected predictors, however predicted probabilities were obtained in different ways. Finally, the same data was used to create a statistically derived prognostic model to which both expert-based models were compared.

**Expert-based model dichotomous**
In this model the ten predictors from our Delphi procedure were included as dichotomous predictors by using their median score as the split value. Subsequently regression coefficient estimates were derived by performing a multivariable logistic regression analysis.

**Expert-based model continuous**
For this model the same steps were conducted as for the derivation of expert model dichotomous. As dichotomizing can lead to information loss,21,22 expert-based selected predictors were included as continuous predictors where possible. Only when predictors failed the linearity assumption predictors were categorised into three groups.

**Statistical model**
In the statistically derived model, predictors were, in contrast to the expert-based models, selected based on significance of the association with the outcome (persistent shoulder pain at 3 months). This selection was preceded by checking the linearity assumption for all potential predictors and if necessary categorisation (three groups) or dichotomization of potential predictors was performed. Furthermore, variable selection was also preceded by checking for
(multi)collinearity (Pearsons r). Since correlated variables may disturb predictor selection,\textsuperscript{23} in case of correlated variables (r ≥0.5) the most easily obtainable variable in clinical practice was obtained for further analyses. Because of the great number of variables, a univariable pre-selection (α≤ .157, a significance level which is comparable with Akaike’s Information Criterion\textsuperscript{24}) was performed to reduce the number of variables. Subsequently, a multivariable regression analysis in combination with a backward selection strategy (α= .157), was performed to obtain the final model.\textsuperscript{13}

Since the derivation of a prognostic model can seriously be affected by missing data,\textsuperscript{25} we used a multiple imputation procedure (MICE, Multivariate Imputation by Chained Equations\textsuperscript{26}) to impute missing values. All models were derived and tested for performance using imputed data. Predictive performance was determined by how well predicted and observed probabilities agreed (calibration indicated by the slope of the calibration plot), how well the models distinguished subjects with and without persistent symptoms (discrimination indicated by the Area Under the receiver operating Curve (AUC) for dichotomous outcomes), explained variance (indicated by Nagelkerke’s R\textsuperscript{2}) and a bootstrap estimation of how much the model performance will deteriorate when applied to new subjects (overoptimism).\textsuperscript{27}

**Software**

A web-based questionnaire was used in order to perform the Delphi procedure (Examine software\textsuperscript{28}). Model derivation and assessment of model performance was performed by using R software version 2.6.0. Logistic regression and the bootstrap internal validation procedure were performed in R by using the R-Design package.
RESULTS

Expert panel

52 Dutch and British health care professionals involved in the assessment and management of (patients with) shoulder disorders were invited to participate in our expert panel. From these 41 (79%) agreed to participate. The self reported primary professions indicated that among the participants were 16 (39%) general practitioners, 16 (39%) physiotherapists, 3 (8%) rheumatologists, 3 (8%) epidemiologists, 1 (3%) manual therapist and 1 (3%) senior lecturer in occupational medicine with a background as a GP and in occupational medicine. Half of all the participants combined their primary profession with a second vocation. From the participating physiotherapists 5 were also certified manual therapists and 3 GP’s were professionally involved in musculoskeletal research. On average, the panel members had 17 (minimum of 5 and a maximum of 35) years of professional experience. Our international panel consisted of 25 (61%) British and 16 (39%) Dutch members. Figure 1 shows that participation of panel members varied from 88% to 82% in the separate Delphi rounds. 29 (71%) panel members completed all three Delphi rounds, 5 (12%) completed two rounds and 3 (7%) dropped out. All panel members contributing to the Delphi study are named in the acknowledgements.

Delphi procedure

First round

In the first Delphi round we provided the expert panel with 46 potential predictors, which the experts ranked according to their predictive abilities. Table 1 presents the mean scores for predictive importance based on round 1. The highest scores were assigned to the variables symptom duration, symptom history, and pain
**Table 1. Delphi procedure results**

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Round 1</th>
<th></th>
<th>Round 2</th>
<th></th>
<th>Round 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>(SD)</td>
<td>rank</td>
<td>mean</td>
<td>rank</td>
<td>inclusion</td>
</tr>
<tr>
<td>symptom duration</td>
<td>4.26</td>
<td>(0.78)</td>
<td>1</td>
<td>253</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>pain catastrophizing</td>
<td>4.03</td>
<td>(0.95)</td>
<td>3</td>
<td>160</td>
<td>2</td>
<td>84%</td>
</tr>
<tr>
<td>symptom history</td>
<td>4.06</td>
<td>(0.80)</td>
<td>2</td>
<td>95</td>
<td>3</td>
<td>90%</td>
</tr>
<tr>
<td>fear-avoidance beliefs</td>
<td>3.91</td>
<td>(1.04)</td>
<td>6</td>
<td>86</td>
<td>4</td>
<td>90%</td>
</tr>
<tr>
<td>coexisting neck pain</td>
<td>3.49</td>
<td>(1.12)</td>
<td>22</td>
<td>79</td>
<td>5</td>
<td>100%</td>
</tr>
<tr>
<td>baseline severity of shoulder disability</td>
<td>3.49</td>
<td>(1.07)</td>
<td>21</td>
<td>78</td>
<td>6</td>
<td>97%</td>
</tr>
<tr>
<td>coexisting psychological complaints / general mental health problems</td>
<td>3.83</td>
<td>(0.95)</td>
<td>9</td>
<td>73</td>
<td>7</td>
<td>84%</td>
</tr>
<tr>
<td>currently on sick leave because of shoulder pain</td>
<td>3.69</td>
<td>(0.87)</td>
<td>11</td>
<td>70</td>
<td>8</td>
<td>81%</td>
</tr>
<tr>
<td>multisite pain</td>
<td>n=10a</td>
<td>-</td>
<td>68</td>
<td>9</td>
<td>84%</td>
<td>7</td>
</tr>
<tr>
<td>somatisation</td>
<td>4.00</td>
<td>(0.97)</td>
<td>4</td>
<td>67</td>
<td>10</td>
<td>84%</td>
</tr>
<tr>
<td>age</td>
<td>3.58</td>
<td>(0.73)</td>
<td>17</td>
<td>64</td>
<td>11</td>
<td>19%</td>
</tr>
</tbody>
</table>
only the 20 most highly ranked predictors from round 2 are ordered by their prognostic importance as assessed in the second Delphi round

<table>
<thead>
<tr>
<th>predictors</th>
<th>mean (SD)</th>
<th>rank</th>
<th>score</th>
<th>consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>shoulder pain intensity</td>
<td>3.49 (1.24)</td>
<td>23</td>
<td>64</td>
<td>13%</td>
</tr>
<tr>
<td>illness perceptions</td>
<td>3.91 (1.01)</td>
<td>5</td>
<td>62</td>
<td>13%</td>
</tr>
<tr>
<td>depression / depressive symptoms</td>
<td>3.91 (1.07)</td>
<td>7</td>
<td>61</td>
<td>10%</td>
</tr>
<tr>
<td>passive coping strategies</td>
<td>3.63 (1.06)</td>
<td>15</td>
<td>60</td>
<td>6%</td>
</tr>
<tr>
<td>repetitive movements</td>
<td>3.86 (0.81)</td>
<td>8</td>
<td>46</td>
<td>3%</td>
</tr>
<tr>
<td>high physical load at work or leisure time</td>
<td>n=11a</td>
<td>-</td>
<td>39</td>
<td>6%</td>
</tr>
<tr>
<td>strain or overuse due to usual activities in work or leisure time</td>
<td>3.43 (1.06)</td>
<td>25</td>
<td>32</td>
<td>10%</td>
</tr>
<tr>
<td>patient reports stiffness of the shoulder</td>
<td>3.06 (0.97)</td>
<td>30</td>
<td>32</td>
<td>6%</td>
</tr>
<tr>
<td>diabetes mellitus</td>
<td>n=2a</td>
<td>-</td>
<td>32</td>
<td>10%</td>
</tr>
</tbody>
</table>

mean: mean score for rating individual potential predictor importance, [1 (not at all predictive) to 5 (highly predictive)]

SD: standard deviation

rank: the order of appearance of potential predictors in individual Delphi rounds based on scores for prognostic importance

sum score: the sum of all scores for predictive importance for each potential important predictors for persistent shoulder pain intensity

predictor inclusion agreement: consensus on the inclusion of potential predictors in the selection of the ten most important predictors for persistent shoulder pain intensity

a: additional potential predictor added by n panel members

b: potential predictor omitted from final selection since it overlaps with ‘fear-avoidance’ and the item is ‘too general’

c: potential predictor combined with predictor ‘severity of shoulder disability’

d: potential predictor omitted from final selection since it overlaps with ‘multisite pain’ and ‘fear avoidance’
catastrophizing with mean (SD) score of respectively 4.3 (0.8), 4.1 (0.8) and 4.0 (0.9) as can be seen in Table 1. Mean scores (SD) for predictive importance ranged from 4.3 (0.8) for symptom duration to 2.0 (0.9) for coexisting knee pain or symptoms. Panel members suggested 19 additional potential predictors. These were added to the 46 listed potential predictors and separately arranged in the order of times they were suggested by individual panel members. Variables mentioned the most were (un)employment (mentioned by 12 panel members), high physical load at work or leisure time (mentioned by 11 panel members) and multisite pain (mentioned by 10 panel members).

**Second round**
The list of potential predictors and additional variables from the first round was re-evaluated in the second Delphi round. Table 1 presents the ranking and sumscores based on round 2 for the 20 predictors with the highest rankings. The results show that two potential predictors were considered to be very important by the panel; symptom duration and pain catastrophizing. Some predictors that were indicated as being moderately predictive in the first
Delphi round (mean score approximately 3.5), were re-evaluated as being more important in the second round (i.e., baseline severity of shoulder disability and coexisting neck pain). On the contrary, the predictors illness perceptions, depression and repetitive movements were in the second round re-evaluated as being of lesser importance than indicated in the first round. From the newly suggested variables only multisite pain was regarded as being highly predictive. Other newly suggested variables were not included in the selection of key predictors. Based on the second round the 10 most important predictors were: symptom duration, pain catastrophizing, symptom history, fear-avoidance, coexisting neck pain, baseline severity of shoulder disability, coexisting psychological symptoms, sick leave because of shoulder pain, multisite pain, somatisation.

Third round

When presented with the ten key predictors as indicated by the second Delphi round, 13 (42%) panel members agreed on this selection. The majority of the experts (58%) however, disagreed with these predictors being the ten most important predictors of persistent shoulder pain. There was uncertainty regarding 5 predictors (pain catastrophizing, coexisting psychological symptoms, sick leave because of shoulder pain, multisite pain, and somatisation) which, as can be seen in Table 1, were selected among the top 10 predictors by \( \leq 85\% \) of the panel members. The main reason for this disagreement was that these predictors were believed to overlap with other included predictors. For instance, the predictor coexisting psychological symptoms was said to overlap with the predictors pain catastrophizing and fear-avoidance; somatisation with multisite pain; and sick leave with baseline shoulder disability. To resolve this problem of overlapping, panel members provided replacement options, such as age, shoulder pain intensity and illness perceptions mentioned by respectively 6, 4 and 4 panel members.
Table 2. Regression coefficients and odds ratios for both the dichotomous and expert-based category

<table>
<thead>
<tr>
<th>Predictors</th>
<th>expert-based categorya</th>
</tr>
</thead>
<tbody>
<tr>
<td>symptom duration</td>
<td>&gt;11 weeks</td>
</tr>
<tr>
<td>presence of the current shoulder pain problem for a period of</td>
<td></td>
</tr>
<tr>
<td>pain catastrophizing</td>
<td>NRS [0-10] &gt;4</td>
</tr>
<tr>
<td>believing shoulder pain to be permanent rather than temporary</td>
<td></td>
</tr>
<tr>
<td>symptom history</td>
<td>yes</td>
</tr>
<tr>
<td>experienced earlier episode(s) of shoulder pain</td>
<td></td>
</tr>
<tr>
<td>fear-avoidance beliefs</td>
<td>NRS [0-10] &gt;7</td>
</tr>
<tr>
<td>believing activity will worsen the shoulder pain</td>
<td></td>
</tr>
<tr>
<td>coexisting neck pain</td>
<td>yes</td>
</tr>
<tr>
<td>additional neck pain during the current shoulder pain period</td>
<td></td>
</tr>
<tr>
<td>severity of shoulder disability</td>
<td>yes</td>
</tr>
<tr>
<td>being unable to perform normal daily activities in the past week</td>
<td></td>
</tr>
<tr>
<td>or for a longer period of time</td>
<td></td>
</tr>
<tr>
<td>multisite pain</td>
<td>yes</td>
</tr>
<tr>
<td>pain or stiffness in other areas than the afflicted shoulder</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>&gt;52 years</td>
</tr>
<tr>
<td>shoulder pain intensity</td>
<td>NRS [0-10] &gt;5</td>
</tr>
<tr>
<td>shoulder pain experienced in the last 24 hours</td>
<td></td>
</tr>
<tr>
<td>illness perceptions</td>
<td>yes</td>
</tr>
<tr>
<td>believing there is not a lot the person can do to control the</td>
<td></td>
</tr>
<tr>
<td>shoulder pain</td>
<td></td>
</tr>
<tr>
<td>intercept</td>
<td></td>
</tr>
</tbody>
</table>

\( \beta \) regression coefficient estimate  
SE standard error of regression coefficient estimate  
OR odds ratio  
95% CI 95% confidence interval for the odds ratio  
a predictors were dichotomized by using median split value scores  
b reference category
prediction in general practice

A continuous expert-based prognostic model for persistent shoulder pain

<table>
<thead>
<tr>
<th>model, dichotomous predictors</th>
<th>expert-based model, continuous predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>category</td>
<td>β (SE) OR (95% CI)</td>
</tr>
<tr>
<td>&lt;6 weeks^b</td>
<td>0.654 (0.183) 1.92 (1.34-2.75)</td>
</tr>
<tr>
<td>&gt;11 weeks</td>
<td>0.549 (0.184) 1.73 (1.21-2.48)</td>
</tr>
<tr>
<td>NRS [0-10]</td>
<td>0.188 (0.181) 1.21 (0.85-1.72)</td>
</tr>
<tr>
<td>NRS [0-10]</td>
<td>-0.031 (0.180) 0.97 (0.68-1.38)</td>
</tr>
<tr>
<td>Score [0-20]</td>
<td>-0.067 (0.207) 0.93 (0.62-1.40)</td>
</tr>
<tr>
<td>years</td>
<td>0.130 (0.178) 1.14 (0.80-1.61)</td>
</tr>
<tr>
<td>NRS [0-10]</td>
<td>0.294 (0.219) 1.34 (0.87-2.06)</td>
</tr>
<tr>
<td>yes</td>
<td>-0.388 (0.186) 0.68 (0.47-0.98)</td>
</tr>
<tr>
<td>yes</td>
<td>0.144 (0.177) 1.15 (0.81-1.63)</td>
</tr>
<tr>
<td></td>
<td>-1.078 (0.226)</td>
</tr>
</tbody>
</table>
These results lead to a new selection of most important predictors which is shown in Table 1. Consensus on this selection was achieved after the third Delphi round. This final set of predictors for persistent shoulder pain three months after initial consultation in primary care, which was agreed on by 29 (97%) panel members (i.e., higher than our predetermined consensus threshold of 90%) included: symptom duration, pain catastrophizing, symptom history, fear-avoidance, coexisting neck pain, severity of shoulder disability, multisite pain, age, shoulder pain intensity and illness perceptions. How these predictors formed the expert-based dichotomous model and the expert-based continuous model can be seen in Table 2.

**Statistically derived model**

As can be seen in Table 3 the statistically derived model included

<table>
<thead>
<tr>
<th>Predictors</th>
<th>category</th>
<th>B</th>
<th>SE</th>
<th>OR</th>
<th>(95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sporting injury</td>
<td>yes</td>
<td>-1.2</td>
<td>(0.499)</td>
<td>0.29</td>
<td>(0.11-0.79)</td>
</tr>
<tr>
<td>longer symptom duration</td>
<td>&lt;6 weeks</td>
<td>0.51</td>
<td>(0.253)</td>
<td>1.67</td>
<td>(1.01-2.75)</td>
</tr>
<tr>
<td></td>
<td>6-11 weeks</td>
<td>0.92</td>
<td>(0.230)</td>
<td>2.51</td>
<td>(1.60-3.96)</td>
</tr>
<tr>
<td></td>
<td>&gt;11 weeks</td>
<td>0.91</td>
<td>(0.233)</td>
<td>2.50</td>
<td>(1.57-3.96)</td>
</tr>
<tr>
<td>coexisting lower back</td>
<td>yes</td>
<td>0.70</td>
<td>(0.298)</td>
<td>2.03</td>
<td>(1.12-3.65)</td>
</tr>
<tr>
<td>inability to perform daily</td>
<td>0 days</td>
<td>-0.5</td>
<td>(0.220)</td>
<td>0.57</td>
<td>(0.37-0.98)</td>
</tr>
<tr>
<td></td>
<td>1-30 days</td>
<td>-0.4</td>
<td>(0.342)</td>
<td>0.65</td>
<td>(0.33-1.29)</td>
</tr>
<tr>
<td></td>
<td>1-12 months</td>
<td>-0.4</td>
<td>(0.342)</td>
<td>0.65</td>
<td>(0.33-1.29)</td>
</tr>
<tr>
<td>coexisting upper</td>
<td>yes</td>
<td>0.34</td>
<td>(0.204)</td>
<td>1.40</td>
<td>(0.94-2.10)</td>
</tr>
<tr>
<td>intercept</td>
<td></td>
<td>-0.7</td>
<td>(0.201)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3. Regression coefficients and odds ratios for the statistical prognostic model for persistent shoulder pain**

B: regression coefficient estimate  
SE: standard error of regression coefficient estimate  
OR: odds ratio  
95% CI: 95% confidence interval for the odds ratio  
a: reference category
the following predictors: sporting injury, symptom duration, coexisting lower back pain, bilateral shoulder pain, inability to perform daily activities, and coexisting upper extremity pain. Only two of these were included in the expert-based selected predictors: symptom duration and baseline severity of the shoulder disability (included in the statistical model as the inability to perform daily activities). However, other statistically selected predictors seem to reflect the expert selected predictor multisite pain (i.e., coexisting back pain, bilateral shoulder pain and upper extremity pain).

How well do our models predict at 3 months

Figure 2 shows the agreement between observed and predicted probabilities for both the statistical and expert-based models in a calibration plot. Following application in the dataset Figure 2 showed that the predicted probabilities for the expert model with dichotomous predictors ranged from 0.18 to 0.76 and from 0.15 to

<table>
<thead>
<tr>
<th>Table 4. Performance measures for the expert-based and statistically derived prognostic models</th>
</tr>
</thead>
<tbody>
<tr>
<td>expert model dicho</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>calibration slope</td>
</tr>
<tr>
<td>R²N</td>
</tr>
<tr>
<td>AUC</td>
</tr>
<tr>
<td>95% CI</td>
</tr>
<tr>
<td>Optimism</td>
</tr>
<tr>
<td>AUC&lt;sub&gt;corrected&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

expert model dicho: prognostic model using dichotomized expert selected predictors
expert model cont: prognostic model using continuous expert selected predictors
R²N: Nagelkerke’s explained variance
AUC: discriminative ability as indicated by the Area Under the ROC Curve
95% CI: 95% confidence interval for the AUC
optimism: the models’ estimated deterioration when applied to new subjects
AUC<sub>corrected</sub>: the optimism corrected Area Under the ROC Curve
0.84 for the expert model with continuous predictors ranged from 0.18 to 0.76 and from 0.15 to 0.84 for the expert model with continuous predictors. The statistically derived model showed a similar range with predicted probabilities between 0.12 and 0.89. Since each model was fitted on the derivation data by multivariable regression analysis, all showed good calibration with calibration points close to the diagonal (i.e., optimal calibration). No differences in calibration slopes between statistical (1.019) and expert-based (1.017 dichotomous and 0.986 continuous) models were observed (Table 4). Table 4 also shows that the expert-based models had lower discriminative abilities compared to the statistical model (AUC expert dichotomous model= 0.656, AUC expert continuous model= 0.679 and AUC statistical model= 0.702). Hence the statistical model distinguishes better between subjects with and without persistent

**Figure 2.** Calibration plot showing the observed versus the predicted probabilities for both the expert-based (dichotomous model: closed black dots, continuous model: open squares) and statistically derived model (closed grey diamonds) for the prognosis of persistent shoulder pain
shoulder pain than the expert-based models. Although the calibration plot showed all models to be well fitted on the derivation data, the internal validation routine showed that regression coefficient estimates of all three models were overoptimistic. This means that when applied in new subjects, predictive performance of all three models are expected to deteriorate.\textsuperscript{25} With bootstrap estimated rates for overoptimism of 0.029 and 0.028, the regression coefficient estimates of the expert based models appeared to be twice as optimistic as the regression coefficient estimates of the statistical model (with an estimated overoptimism of 0.014). Therefore the statistical model is expected to perform better when applied in new subjects.\textsuperscript{27}

**DISCUSSION**

This Delphi procedure resulted in professional consensus on the 10 most important predictors of persistent shoulder pain 3 months after initial consultation in primary care. Expert selected predictors appeared to be different from that of a statistically derived model, however both models confirmed the importance of symptom duration, baseline level of disability and multisite pain. Panel members additionally selected age, baseline pain intensity and psychological factors as important predictors. Concerning predictive performance, we found the statistically derived model to be slightly better than the expert-based prognostic model.

Since clinical expertise is expected to complement statistically derived prognostic models, this study aimed to reach clinical consensus on which predictors are most important for predicting persistent shoulder pain. It was shown that health care professionals’ consensus based selection of key predictors reflected
most statistically selected predictors but also included additional predictors which were not identified by statistical selection. During the inventory of potential predictors (i.e., the first Delphi round) health care professionals even identified predictors which previously have not been directly associated with shoulder pain together with predictors which have been shown to be associated with poor outcome in other musculoskeletal pain conditions (e.g., earlier experiences with shoulder treatment, smoking, diabetes mellitus, alcohol intake, ethnicity, level of training discipline, perceived versus actual work activity, social support, distress). None of these predictors made it to the final selection of most important predictors.

The consensus based selection of the key predictors of persistent shoulder pain, was derived using a Delphi procedure. Although this technique is often applied in consensus based research, its validity and reliability are sometimes object of discussion. Since consensus findings may vary depending on the panel, the guidelines for consensus methods by Fink et al. were followed where possible. With a minimum participation rate of 31 panel members in a single Delphi round, our expert panel was sufficiently sized for obtaining reliable results. As multi-disciplinary panels may select a wider range of predictors compared to single-disciplinary panels, our panel consisted of health care professionals and researchers from different disciplines and geographical areas in the United Kingdom and the Netherlands. Furthermore, the Delphi procedure was completely anonymous. Panel members never met, neither did they know each others’ identities. Therefore, negative group interactions or dominant opinions were eliminated. To assist our panel members in selecting prognostic factors we provided them with a resource, i.e., a list of potential predictors based on a previous systematic review. Although not an uncommon practice in consensus based research,
one might argue that providing such a list might hinder the unveiling of new potential predictors. Therefore, during the entire Delphi process all panel members were encouraged to suggest additional potential predictors. Since a part of our panel was also involved in shoulder related clinical research, they were considered to be informed on the latest developments in the literature. This together with the option of providing additional information lead us to believe all predictors for persistent shoulder pain in primary care patients were identified by our panel.

How can we explain observed differences in expert and statistical selected prognostic factors? Taking into account the above mentioned considerations, it is unlikely that these differences were caused by methodological limitations in the Delphi procedure. Because our panel of health care professionals were trained in the clinical management of individual patients, they might have had problems with providing prognostic factors for the general population of shoulder pain patients. This could have complicated the identification of universal prognostic factors for shoulder pain patients.

Another explanation for the observed differences in selected predictors might be found in the methodological limitations of predictor selection in statistically derived models. In the applied methodology, predictors were selected by an automated selection procedure. As shown by Austin and Tu,\textsuperscript{2} statistical predictor selection can give biased results. Automated backward elimination or forward selection might result in omission of important predictors or the random selection of less important predictors. As a result statistically derived models may be unstable,\textsuperscript{2} which was previously demonstrated for our statistically derived model.\textsuperscript{13} Differences between expert-based and statistical selection of predictors might therefore be largely influenced by the chosen method of statistical
predictor selection. However, how to optimally perform variable selection is still a subject of discussion.\textsuperscript{35}

**Strengths and weaknesses**

One of the strengths of the current study was that next to establishing consensus on key predictors, the predictive performance of these predictors was empirically tested. Results showed that both expert-based models did not perform as optimally compared to the statistically derived prognostic model. This is a notable result since clinical knowledge is expected to complement statistical modelling and the derivation of our statistical model has some known limitations in predictor selection. These findings do however need to be interpreted with caution since they do not suggest that statistical based scoring systems are superior to clinical prognosis. Although we asked our panel for suggestions on how to formulate and score each predictor, a weakness of this study was that we had to use an existing dataset which did not include the exact same variables as proposed by the expert panel.

Another weakness was that a potential floor-effect associated with low baseline pain ratings could have occurred in our measure of outcome. Although approximately 19\% of the subjects in our database had a baseline pain score of $\leq$2, all baseline pain categories (e.g., 1 to 10) showed a constant percentage of subjects identified with persistent shoulder pain of approximately 40 to 60\%. Thus, apart from subjects with a baseline pain rating of 0 we reasoned that our analyses were not affected by a potential floor effect. Furthermore, although we derived an optimal model using continuous scales, the expert-based model had to compete with a statistical model that was derived in the same dataset and therefore was expected to show better predictive performance. Hence the
conclusion of the superiority of statistical prognosis over clinical prognosis might be impetuous.

Another aspect that can be regarded as a weakness of our study is the dichotomization of key predictors in one of the expert-based prognostic models. Dichotomization of predictors is in the literature often criticized because it may lead to loss of information and thus a decrease in predictive performance.\textsuperscript{22} Although we expected our panel members to be familiar with this undesirable effect, most of them said they preferred a prognostic model which consists of simple (i.e., dichotomous) predictors. This illustrates at this point the discrepancy between prognostic research and clinical practice. In prognostic research model performance is most important, in clinical practice models in addition need to be easy to use. Unfortunately simplicity of the model goes at cost of the predictive performance, as can be seen by the effect of dichotomisation of predictors by using median values as cut-off.\textsuperscript{21,22}

With these considerations it remains unclear whether estimations of prognosis by health care professionals is superior or not to the estimation of prognosis obtained by scoring systems. Previous studies have shown that both clinical prognosis and scoring systems can be superior to one another.\textsuperscript{36-38} It might even be conceivable that prognostic superiority is case dependent (type of musculoskeletal condition, health care profession). Therefore, clinical prognosis and scoring systems for the prognosis of non-recovery from shoulder pain will be compared in a future study.

\textbf{Conclusions}

As clinical expertise is expected to complement statistically based prognostic research we showed that an expert panel of health care professionals and researchers indeed selected additional
predictors compared to a statistical selection procedure, emphasizing the expected value of age, shoulder pain intensity and psychological factors. Both selections confirmed the importance of symptom duration, baseline severity of shoulder disability and multisite pain in the prognosis of shoulder pain. When transformed to a prognostic model, the expert-based models performed less well compared to a statistically derived model. Application in a different population than the derivation dataset should demonstrate whether statistically derived models are indeed less overoptimistic compared to expert-based models.
References


PREDICTION MODELS IN GENERAL PRACTICE

the value of prognostic models for persistent shoulder pain in general practice

ABSTRACT

Objectives:Recently we have derived several prognostic models for the persistence of shoulder pain three months after initial consultation in general practice. Different approaches were used to derive these models using either statistical methods or clinical expertise to select prognostic indicators. The objective of this study was to investigate the external validity of the two most promising models and assess their value in general practice by comparing their predictive performance to the general practitioners’ own prognosis in routine clinical practice.

Methods:In a validation cohort of 203 patients, predictive performance of the models was assessed, and performance measures (discrimination, addressed by the Area Under the receiver operator characteristic Curve, AUC, and the calibration) were compared to the general practitioners’ estimation of the risk of persistent shoulder pain three months after consultation.

Results:Both prognostic models performed suboptimal with AUC values of 0.580 and 0.625 for the statistical and expert-based models respectively. This performance was only slightly better than the general practitioners’ clinical prognosis which showed an AUC of 0.506. The prognostic models were shown to overestimate the risk of persistent shoulder pain whereas the general practitioners underestimated this risk.

Conclusion:This study suggests that prognostic models for persistent shoulder pain could potentially be useful in primary care. Although the evaluated models showed inadequate performance and would need to be improved.
INTRODUCTION

With a one-year prevalence of approximately 31%\textsuperscript{1} shoulder pain is a commonly occurring musculoskeletal complaint in the Dutch general population. The incidence in Dutch general practice is estimated at 58 consultations per 1000 consulting patients each year.\textsuperscript{2} The clinical course is unfavourable in many patients with recovery levels of 30\%, 50\% and 60\% at six weeks, six months and one year respectively.\textsuperscript{3-5} Early identification of patients at risk of persistent symptoms may enable timely intervention by clinicians and thus, at least in some cases, may prevent long term activity limitations, work absence and high healthcare resource use.\textsuperscript{5,6}

The application of clinical prediction models for patients with shoulder complaints\textsuperscript{7,8} may help clinicians to estimate the likely future outcome of the condition in individual patients. By using early signs or symptoms and patient characteristics these models generate a quantitative estimate of the absolute risk of persistent complaints (or pain) for individual patients. Information regarding which patients are likely to develop persistent complaints and who are likely to recover in the near future may subsequently be used by clinicians to make decisions regarding the need for treatment or referral, and to inform patients of their prognosis.

Prediction models typically perform well in patients whose data were used for model derivation. When applied in other patient populations however, the clinical performance often deteriorates.\textsuperscript{9} This declining performance may be caused by differences in patient characteristics between the derivation sample and subsequent patient samples\textsuperscript{10,11} or from methodological deficiencies in the model building process.\textsuperscript{12}

In order to examine whether a newly derived prediction model can be meaningfully used in clinical practice, first its
performance in similar patients needs to be tested and found acceptable.\textsuperscript{11,13} This is called internal validation.\textsuperscript{14,15} Next, model performance should be tested prospectively in other patient populations representing the setting and population for which the model has been designed. This external validation is most important to examine the models’ generalizability.\textsuperscript{9}

Next to the models’ validity, their addition to a clinicians’ unaided prognostic estimate is also of great importance when considering clinical utility or usefulness.\textsuperscript{11} Obviously, a prediction model that is not better than a clinician’s own estimate of future outcome has little benefit in clinical practice. Therefore, to determine the clinical usefulness of a predictive model its prognostic performance needs to be compared to the performance of the clinician’s own prognostic estimate.

Recently the authors have derived several prognostic models for the prediction of persistent shoulder pain three months after initial consultation in general practice.\textsuperscript{16-18} These models differed in the way they were derived: with little or much clinical input,\textsuperscript{16} by recursive partitioning\textsuperscript{18} or with various statistical methods to account for missing data.\textsuperscript{17} Internal validation of these models has already been undertaken.\textsuperscript{16-18} Their external validity and the added value in clinical practice still needs to be examined. The two most promising models in terms of performance and internal validity were selected for external validation.

The aim of this study is twofold, firstly to investigate the external validity of the two previously derived models in a new prospective sample of patients consulting their general practitioner with a new episode of shoulder complaints, and secondly to determine to compare their prognostic estimates with the clinicians’ own assessment of the likely outcome of the shoulder pain problem in
routine clinical practice.

METHODS

Patients

An external validation cohort (Shoulder Pain Prognosis, SPP) was formed based on patients presenting with a new episode of shoulder pain in Dutch general practice. Since a minimum sample size of 100 events and 100 non-events has been suggested in order to detect substantial changes in accuracy, and the incidence of persistent shoulder pain in the derivation cohort was around 50% we set out to gather a cohort of 200 patients with a new episode of shoulder pain. Where possible, similar selection criteria and data acquisition procedures were used as in the previous Dutch Shoulder Study (DSS), which was the derivation cohort for the previously developed prediction models. Between August 2009 and January 2012, 35 general practitioners (GPs) in 23 family practices recruited patients at first consultation for a new episode of shoulder pain in two geographical areas in the Netherlands (Amsterdam, Groningen). From these, 7 practices previously participated in the DSS study. Next to recruitment by GPs during consultation, the electronic databases of the practices were periodically searched retrospectively for eligible patients not yet invited to participate. Included were patients who were older than 18 years, had not consulted their GP in the preceding three months for the afflicted shoulder, had no cognitive limitations which could hinder completion of written questionnaires, and had no severe physical or psychological conditions such as fractures or dislocations in the shoulder region, rheumatic disease, neoplasms, neurological or vascular disorders or dementia. All eligible patients received a written invitation to participate and provided written informed consent. The validation
study was approved by the Medical Ethics Committee of the VU University Medical Center, Amsterdam, The Netherlands.

**Prognostic models**

Outcome data were collected by postal questionnaire at three months follow-up. Persistent shoulder pain was defined as improvement in shoulder pain intensity (scored by participants on a 0-10 numeric rating scale) of 50% or less over three months. The following two previously developed prognostic models for predicting persistent shoulder pain\textsuperscript{16,17} were tested for external validity.

**Statistical model**

This model was developed based on data from the DSS derivation cohort using the methodology for creating a prediction model as described by Harrell et al.\textsuperscript{12} Candidate predictors were obtained using information from a systematic review of prognostic studies in shoulder pain\textsuperscript{20}, and an optimal set of predictors was statistically selected using a multivariable logistic regression analysis with an automated backward variable selection procedure (critical P-value of 0.157). To account for missing values in the derivation cohort which could seriously affect modelling results,\textsuperscript{21} a multiple imputation technique (MICE, Multivariate Imputation by Chained Equations\textsuperscript{22}) was applied. This led to the derivation of a prognostic model consisting of 6 predictors: sporting injury, longer duration of complaints, concomitant lower back pain, involvement of both shoulders, inability to perform daily activities and concomitant upper extremity pain. For more details see the persistent pain intensity MI·5 model in Vergouw et al.\textsuperscript{16}

**Expert-based model**

This model was also derived using a multivariable logistic regression
analysis however predictor selection was based on a Delphi study among clinical experts. An international group of 41 mainly health care professionals involved in the assessment and management of shoulder pain reached consensus in three subsequent Delphi rounds on 10 key predictors for persistent shoulder pain. Using the DSS cohort (missing values were imputed using MI) the scores on key predictors were extracted, using continuous measures where possible. Multivariable logistic regression analysis was used to derive weights, i.e., regression coefficient estimates, for each predictor and develop a prediction model based on clinical expertise. The 10 included predictors were: longer symptom duration, higher pain catastrophizing, presence of widespread symptoms, more fear-avoidance beliefs, coexisting neck pain, higher severity of shoulder disability at baseline, multisite pain, higher age, higher pain intensity, and more illness perceptions. For more details see the expert-based continuous model in Vergouw et al.

**General practitioners’ prognosis**

In order to compare the prognostic models to prognosis in clinical practice (clinical usefulness), each GP provided their own estimate of the likelihood of persistent shoulder pain at three months for each participant immediately after initial consultation. GPs were asked to score their estimate of the probability of persistent shoulder pain at 3 months on a scale from 0-100% indicating being (very) certain that pain would persist. None of the participating GPs was familiar with the content of any of the prognostic models tested.

**Analysis**

Both models were applied using the original regression coefficients as estimated in the derivation cohort, and with the original regression coefficients corrected for overoptimism by a bootstrapping
Table 1. *Baseline characteristics of the validation (SPP, n=203) cohort and the derivation (DSS, n=587) cohort.*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
<th>Validation SPP (n=203)</th>
<th>Derivation DSS (n=587)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>male (vs. female)</td>
<td>53%</td>
<td>50%</td>
</tr>
<tr>
<td>age (years)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>mean (SD)</td>
<td>53 (13.5)</td>
<td>51 (14.0)</td>
</tr>
<tr>
<td>education</td>
<td>low</td>
<td>37%</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>middle</td>
<td>36%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>27%</td>
<td>23%</td>
</tr>
<tr>
<td>employment</td>
<td>employed (vs. unemployed)</td>
<td>61%</td>
<td>60%</td>
</tr>
<tr>
<td><strong>Disease characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>history (life)</td>
<td>experienced earlier episodes in life</td>
<td>60%&lt;sup&gt;*&lt;/sup&gt;</td>
<td>48%</td>
</tr>
<tr>
<td>history (past year)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>experienced earlier episodes in the past year</td>
<td>62%</td>
<td>55%</td>
</tr>
<tr>
<td>symptom duration&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>0-6 weeks</td>
<td>34%</td>
<td>35%</td>
</tr>
<tr>
<td></td>
<td>7-12 weeks</td>
<td>34%&lt;sup&gt;*&lt;/sup&gt;</td>
<td>24%</td>
</tr>
<tr>
<td></td>
<td>&gt;3 months</td>
<td>31%&lt;sup&gt;*&lt;/sup&gt;</td>
<td>41%</td>
</tr>
<tr>
<td>onset</td>
<td>gradual (vs. Acute)</td>
<td>60%</td>
<td>62%</td>
</tr>
<tr>
<td>cause</td>
<td>unexpected load/movement</td>
<td>5%</td>
<td>6%</td>
</tr>
<tr>
<td></td>
<td>strain/overload unusual activities</td>
<td>12%</td>
<td>16%</td>
</tr>
<tr>
<td></td>
<td>strain/overload usual activities</td>
<td>20%&lt;sup&gt;*&lt;/sup&gt;</td>
<td>6%</td>
</tr>
<tr>
<td>accident</td>
<td></td>
<td>5%&lt;sup&gt;*&lt;/sup&gt;</td>
<td>23%</td>
</tr>
<tr>
<td>sporting injury&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td>6%</td>
<td>5%</td>
</tr>
<tr>
<td>unclear</td>
<td></td>
<td>28%&lt;sup&gt;*&lt;/sup&gt;</td>
<td>41%</td>
</tr>
<tr>
<td>other</td>
<td></td>
<td>18%&lt;sup&gt;*&lt;/sup&gt;</td>
<td>10%</td>
</tr>
<tr>
<td>afflicted shoulder&lt;sup&gt;b&lt;/sup&gt;</td>
<td>bilateral (vs. unilateral)</td>
<td>10%</td>
<td>13%</td>
</tr>
<tr>
<td></td>
<td>mean (SD)</td>
<td>mean (SD)</td>
<td>mean (SD)</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>pain intensity [0-10](^a)</td>
<td>4.9 (2.27)</td>
<td>4.8 (2.32)</td>
<td></td>
</tr>
<tr>
<td>SDQ score [0-100]</td>
<td>61 (20.9)</td>
<td>60 (24.2)</td>
<td></td>
</tr>
<tr>
<td>disability [0-20](^a)</td>
<td>12 (4.2)</td>
<td>12 (4.6)</td>
<td></td>
</tr>
<tr>
<td>multisite pain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>co-morbidity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neck pain(^a)</td>
<td>22%*</td>
<td>34%</td>
<td></td>
</tr>
<tr>
<td>high back pain</td>
<td>8%</td>
<td>9%</td>
<td></td>
</tr>
<tr>
<td>low back pain(^b)</td>
<td>19%</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>upper extremity pain(^b)</td>
<td>31%</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>lower extremity pain</td>
<td>27%</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td><strong>Psychological factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCCL catastrophizing</td>
<td>2.3 (0.88)</td>
<td>2.2 (0.83)</td>
<td></td>
</tr>
<tr>
<td>pain catastrophizing(^a)</td>
<td>5.6* (2.98)</td>
<td>3.9 (3.09)</td>
<td></td>
</tr>
<tr>
<td>FABQ score</td>
<td>14.3 (5.4)</td>
<td>14.1 (5.6)</td>
<td></td>
</tr>
<tr>
<td>fear-avoidance beliefs(^a)</td>
<td>6.7 (3.05)</td>
<td>6.2 (3.28)</td>
<td></td>
</tr>
<tr>
<td>illness perceptions(^a)</td>
<td>2.9* (1.32)</td>
<td>3.4 (1.76)</td>
<td></td>
</tr>
<tr>
<td><strong>Physical factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sporting activities</td>
<td>active in sports</td>
<td>56%</td>
<td>54%</td>
</tr>
<tr>
<td>physically active</td>
<td>less than others</td>
<td>18%</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>equally active</td>
<td>47%</td>
<td>42%</td>
</tr>
<tr>
<td></td>
<td>more active</td>
<td>35%</td>
<td>39%</td>
</tr>
<tr>
<td>inability to perform daily activities(^b)</td>
<td>0 days</td>
<td>47%*</td>
<td>58%</td>
</tr>
<tr>
<td></td>
<td>1-30 days</td>
<td>42%*</td>
<td>32%</td>
</tr>
<tr>
<td></td>
<td>1-12 months</td>
<td>10%</td>
<td>10%</td>
</tr>
</tbody>
</table>

\(^a\) predictor in the expert-based prognostic model  
\(^b\) predictor in the statistically derived prognostic model  
* variables with substantial different values in the validation and derivation cohort
technique. To evaluate and compare the prognostic performance of the statistically derived model, expert-based model and the GPs’ prognostic, discrimination and calibration were assessed. Discrimination, or how well a model distinguishes between patients with and without persistent shoulder pain, was quantified by the c-index which for binary outcomes is identical to the area under the receiver operator characteristic curve (AUC). Calibration, or the agreement between estimated prognosis and actual persistent pain intensity, was measured by computing the slope of the calibration plot (estimated prognosis vs. observed frequencies of persistent pain intensity).

**Software**

All analyses were performed using R software (version 2.6.0) with use of the R rms package.

**RESULTS**

**Data**

Between August 2009 and January 2012, 242 patients from 23 family practices in the Netherlands were found eligible for participation in the SPP study. All were, on average 9 days after initial consultation, send a baseline questionnaire which resulted in a response rate of 89% (n=215). Follow-up questionnaires were sent on average at 96 days after initial consultation and 6% (n=12) of the participants were lost to follow-up. As a result the SPP cohort consisted of 203 patients with complete baseline and follow-up data. To compare the characteristics of the derivation and validation cohorts, baseline characteristics of the SPP cohort and the DSS derivation set are described in Table 1. Some differences between

---

126
baseline characteristics of the SPP validation cohort and the DSS derivation set are described in table 1. Some differences between both cohorts were observed, possibly indicating more severe shoulder complaints in the validation cohort.

Outcome
In the SPP cohort, 92 patients (45%) were classified as having persistent shoulder pain intensity after three months of follow-up, which was similar to the percentage of participants with persistent pain observed in the derivation cohort (47%).

Model performance
Table 2 shows the results of applying the statistically derived and expert-based prognostic models in the external validation cohort. Both prognostic models show a decrease in AUC values compared to their internally validated performances in the derivation cohort. For the statistical model the AUC dropped from 0.688 (internally validated performance in the derivation cohort) to 0.580 in the validation cohort. For the expert-based model the AUC dropped from 0.651 (internally validated performance in the derivation cohort) to 0.625 in the validation cohort. The calibration slopes also indicate poor performance of the prognostic models in the validation cohort, as can be seen from figure 1. Calibration slopes were 0.375 for the statistical and 0.757 for the expert-based model.

Clinical usefulness
When predicted probabilities from the (shrunken) prognostic models are compared to the GPs’ prognosis, GPs appear to provide estimates indicating a lower risk of persistent shoulder pain (31.9% vs 68.6% and 61.3% for the GPs’ estimates, expert-based model and statistical
Table 2. Performance measures of the statistical and expert-based prognostic model and the GP prognostic estimates.

<table>
<thead>
<tr>
<th>Shrinkage factor</th>
<th>Statistical model</th>
<th>Expert-based model</th>
<th>GP prognosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>original β’s</td>
<td>shrunken β’s</td>
<td>original β’s</td>
</tr>
<tr>
<td>shrinkage factor</td>
<td>-</td>
<td>0.907</td>
<td>-</td>
</tr>
<tr>
<td>AUC&lt;sub&gt;apparent&lt;/sub&gt;</td>
<td>0.702</td>
<td>0.702</td>
<td>0.679</td>
</tr>
<tr>
<td>(95% CI)</td>
<td>(.660-.745)</td>
<td>(.660-.745)</td>
<td>(.636-.722)</td>
</tr>
<tr>
<td>AUC&lt;sub&gt;optimism corrected&lt;/sub&gt;</td>
<td>0.688</td>
<td>-</td>
<td>0.651</td>
</tr>
<tr>
<td>AUC&lt;sub&gt;external validation&lt;/sub&gt;</td>
<td>0.580</td>
<td>0.580</td>
<td>0.625</td>
</tr>
<tr>
<td>(95% CI)</td>
<td>(.502-.659)</td>
<td>(.502-.659)</td>
<td>(.548-.702)</td>
</tr>
</tbody>
</table>
model respectively). As can be seen from the histograms in the calibration plots (figure 1), the distribution of the GPs’ prognosis is skewed to the right whereas the predicted probabilities from the prognostic models resemble a normal distribution. The calibration plot shows that GPs underestimated and prognostic models overestimated the probability of persistent shoulder pain at three months follow-up. With an AUC of 0.506 the participating GPs were less successful in distinguishing between patients with or without poor outcome compared to the prognostic models (table 2).

**Figure 2.** Calibration plots for the statistical and expert-based prognostic models and for the GP prognosis.
DISCUSSION
The aim of this study was to determine the external validity of a statistically derived and an expert-based prognostic model for persistent shoulder pain, and to compare these with the GPs’ own estimate of the likely outcome of the shoulder pain problem. The results showed that both prognostic models performed suboptimal and only slightly better in terms of discrimination and calibration compared to the GPs’ own prognosis.

External validation of the models for persistent shoulder pain showed a larger decrease in predictive performance despite the use of optimism corrected coefficients following an internal validation procedure in the derivation cohort. This confirms that internal validation is not per se indicative for performance in new patients and therefore is no substitute for external validation. An AUC of around 0.6, as seen in our validation cohort, is often regarded as indicative for poor discriminative ability and therefore not considered to be clinically useful. When the validated models were compared to the GPs’ prognosis based on their own assessment of the patient’s history and physical examination, the prognostic models showed slightly better performance. This suggests that prognostic models for persistent shoulder pain could potentially be useful in primary care. However, the current models showed inadequate performance and would need to be improved. Thus, as demonstrated before at 6 months follow-up, estimating future outcome in patients presenting with new episodes of shoulder pain in primary care might be difficult.

The results regarding external validation of the two different prognostic models are of interest. Where the statistically derived model showed only slight optimism after the internal validation, in an external validation cohort model performance showed a
significant drop (from an AUC of 0.688 to 0.580). The expert-based model however, showed a smaller reduction in performance in the validation cohort (from 0.651 to 0.625). An explanation for this might be found in the different methods of selecting predictors. The expert-based model reflects consensus on important predictors among shoulder related health care professionals, which is based on their experiences of managing shoulder pain in routine clinical practice. While the statistical model was based on statistical predictor selection. This means that the latter model was derived from a large number of candidate predictors by using univariable and multivariable techniques, using a single cohort. Although not uncommon, these methods are often criticized since predictor selection is an important cause of overoptimistic regression models and results are highly dependent on the characteristics of the derivation cohort for which resulting models might be unstable.\textsuperscript{26,27} Since our bootstrap internal validation procedure, a method postulated by Efron,\textsuperscript{28} did not incorporate predictor selection, the overoptimism in the statistical model was therefore presumably underestimated. These problems can be overcome by reducing the number of candidate predictors. However, given limited available evidence on prognostic factors in shoulder pain\textsuperscript{20} it is difficult to prioritise candidate predictors. In addition, the sample size of the derivation cohort was rather small (n=587), which may have complicated finding the optimal model composition. All in all, the ‘best’ way of selecting a multivariable model is still unresolved.\textsuperscript{29}

Next to statistical issues, poor external validation might also be explained by some differences between the derivation and validation cohorts. In our study we aimed to test the generalizability of the derived models over time in similar patients and in a similar primary care setting, focussing on temporal validation. Since this type of validation is subject to some forms of bias, we tried to assess
the risk of such bias. Similar inclusion and exclusion criteria were used for recruiting both the derivation and the validation cohorts, which reduced the risk of selection bias. Furthermore, the data on predictors and outcome measures was collected in the same manner as in the derivation cohort, using postal questionnaires and definitions of predictors and outcome were the same. Therefore the risk of information bias seems low. The risk of persistent shoulder pain was very similar in the derivation and validation cohorts, which seems to indicate the same spectrum of severity of the condition. However, comparison of baseline characteristics showed a number of differences between the two cohorts, possibly indicating more severe shoulder problems in the validation cohort or due to random variation. This might partly explain the drop in performance upon external validation.

The accuracy of the GP estimations of the likely outcome was possibly complicated by the fact that as in the derivation cohort GPs in the validation cohort were not restricted in providing usual care. Therefore, patients with poor prognosis at baseline were more likely to receive more intense treatment (e.g., physiotherapy or specialist referral, injections) compared to patients with better prognosis. When treatment is effective, associations between predictors and outcome will be distorted and the GP prognosis will overestimate the observed outcome. However, our results showed that GPs enrolled in our study underestimated the risk of the likely outcome of the shoulder pain problem. Although GPs possibly based their treatment decisions on initial prognosis, they might have adjusted their risk estimates because they expected treatment to effectively reduce the persistence of shoulder pain. When treatment is less effective than expected, the likely outcome is underestimated as was the case.
Strengths and weaknesses

One of the main strengths of this study is that next to externally validating the derived prognostic models, the performance of these models was compared to the GPs’ own estimate of the likely outcome of the shoulder pain problem. In doing so, this study was, to our knowledge, the first to show that outcome prediction in shoulder pain by GPs is difficult and that our statistically derived models adds little to this.

Although GP confidence in the success of their prescribed treatment might have caused this disappointing prognostication, we cannot completely rule out the role of our outcome definition “improvement of less than 50% in shoulder pain intensity score”. While this threshold was shown to reflect a minimal important change,16 its meaning in clinical practice might be less evident. Instead of estimating future outcome based on decreasing pain scores, it might be more meaningful for GPs to estimate the risk of more apparent outcomes such as overall recovery from shoulder symptoms. Or they might not consider the achieved degree of improvement, but the actual level of pain after 3 months. Furthermore, the GP prognosis may have been influenced by the retrospective database search for eligible patients. By this a time lag between initial consultation and GP’s estimates of approximately 14 days was created for 25% of the included patients. When the estimates of these groups were compared, mean predicted probabilities of persistent shoulder pain for retrospectively included patients were higher than probabilities estimated at consultation (38% vs. 30%). However, the risk of persistent complaints was also higher for retrospectively included patients (56% vs. 42%). Therefore we assume that the retrospective inclusion is unlikely to have influenced GPs’ estimate of outcome.
A final potential weakness of our study is that recruitment was slow, taking 30 months to enrol the required sample size of 200 participants for the validation cohort. Given the incidence of shoulder pain in primary care it is likely that we have missed potentially eligible shoulder patients which might have led to a selective sample. However, if such a selection occurred we do not expect this to be different from the derivation set. Although the baseline characteristics showed some differences between the derivation and validation cohort, the mean number of identified patients per GP was 2.9 for both cohorts (203 patients for 35 GPs in the validation set and 587 patients for 103 GPs in the derivation set), and the risk of persistent shoulder pain was very similar.

**Conclusion**

By comparing two previously derived prognostic models with clinicians’ estimation of the likely outcome of the shoulder pain problem, this study highlighted the difficulties in estimating outcome in primary care patients with shoulder complaints. The 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.
References


7

GENERAL DISCUSSION
GENERAL DISCUSSION

The main objective of this thesis was to study methodological issues that might hinder the derivation of the most valid and best predictive prediction model for persistent shoulder pain in general practice. At the same time, we aimed to assess the predictive performance of different models in a primary care population of patients consulting with a new episode of shoulder pain and compare this to the estimates regarding the prognosis of persistent shoulder pain in individual patients by general practitioners. In this chapter the core findings from this research will be summarized, critically appraised and put in to a broader perspective for both methodological and clinical objectives. Finally we will conclude with some recommendations for future prognostic research and research practice.
Main findings

This thesis is focussed on methodological issues in deriving a valid prognostic model. Since it is known from the literature that several methodological issues influence the results of prognostic modelling, we aimed to illustrate the effects of some of the most important methodological issues: missing data and model instability.

Effects of missing data

Using a cohort of patients with low back pain, we empirically illustrated, how the absence of information due to missing data in baseline characteristics (predictors) and follow-up (outcome) affects the derived prognostic model. We illustrated the superiority of handling missing data by Multiple Imputation (MI) over the commonly used but suboptimal method of Complete Case Analysis (CCA). Model performance, overoptimism and model composition were shown to differ between methods.

Model instability

Next we showed that these differences did also occur when MI was compared with CCA in a cohort of patients with shoulder complaints with missing values in baseline characteristics. Although prognostic modelling in case of missing data benefits from applying multiple imputation, we showed that the models derived for predicting persistent shoulder pain using imputed data sets might still be unstable.

Alternative method: Classification And Regression Tree

Alternative methods for the usually employed regression analysis are sometimes reported in the literature. One of the most commonly reported alternative modelling strategies; the Classification And Regression Tree (CART) methodology, is because of its tree-like
structure, believed to result in more clinically intuitive models. We therefore derived a CART model in a cohort of patients with shoulder pain. Resulting models were however found to be highly overoptimistic and showed poor performance when we adjusted for overoptimism.

**Clinical key predictors of persistent shoulder pain**
Because statistical predictor selection is known to give biased results and the best way of selecting potential predictors is still unresolved, we invited health care professionals in order to take part in a Delphi study in order to capture clinical knowledge on key predictors for persistent shoulder pain. Although some of the key predictors highlighted by clinicians were included in our statistically derived prediction model (e.g., pain duration), we found that clinicians also mentioned key predictors that for which limited prognostic evidence is available in the medical literature (e.g., pain catastrophizing, fear-avoidance beliefs and illness perceptions). With these predictors we developed an expert-based prediction model.

**External validation and value in clinical practice**
When we tested the internal validity of the statistically derived persistent shoulder pain models, which provides an estimation of performance of a prediction model in comparable patients, the model based on multiple imputed data sets showed the best performance (c-statistic 0.688). To determine the models’ generalizability we externally validated their performance by creating a new cohort of patients consulting with a new episode of shoulder pain in primary care. Both the statistical and expert-based models showed poorer performance than estimated in the derivation cohort (c-statistic 0.580 and 0.625 for the statistical and expert-based models respectively). Subsequently, we compared both models to an
estimate of future outcome (probability of persistent pain after three months) provided by the participant’s GPs. It turned out that the GPs underestimated the probability of persistent shoulder pain, and their predictive performance was somewhat lower compared to our developed models. However the derived models overestimated the risk of persistent shoulder pain and were thus still not optimal.

How to appreciate these findings will be discussed in the following sections.

**Missing values and multiple imputation in prognostic research**

The troubling factor of missing data in the development of a prognostic model has recently gained interest in the literature. In chapters 2 and 3 we empirically illustrated the differences in model performance, internal validity and model composition, of using the advocated method of MI to the suboptimal method of CCA in a musculoskeletal prognostic study. Our models derived by CCA performed better in the derivation data set compared to our models that were derived using MI. However, when assessing internal validity in comparable patients the CCA model showed a far greater overoptimism than the MI model. Therefore the MI model performed better than the CCA when performance was corrected for overoptimism. Next, models derived by CCA and MI included other predictors. MI models were therefore regarded as superior over the CCA models. It needs to be addressed however, that due to our empirical approach, knowledge on the “true” predictors of outcome was lacking and in our example results are essentially based on observed differences in performance. However, in addition to our findings regarding model performance, our recommendation to use MI in case of missing data was based on results from previous simulation studies in which MI proved to result in more valid
predictor selection and less biased regression coefficient estimates compared to CCA\textsuperscript{3,4}. Furthermore, our approach illustrated how issues such as the mechanisms responsible for missing values, the number of administered replicate data sets and the combination of multiple imputed data sets into a single result might be addressed empirically in prognostic research.

Determining the reason(s) for missing data is important.\textsuperscript{4-9} Since MI can only use information available in the data set (including information on predictor and outcome variables) to estimate the value of missing observations, MI cannot account for the disturbing influence of missing values when information is missing for unobserved reasons (Missing Not At Random, MNAR).\textsuperscript{6} For MI to be reliably employed, missing values need to occur completely at random (Missing Completely At Random, MCAR) or missingness needs to be associated with other observed patient characteristics (Missing At Random, MAR).\textsuperscript{6} Because it cannot be empirically tested whether data follow the MNAR pattern, we are not completely sure that our MI approach was the most capable method of minimizing bias caused by missing data. However, when the derivation data contains a large number of candidate predictors, the chances of meeting MAR conditions increase. The fact that our derivation set contained a large amount of information and missing values showed associations with other potential predictors led us to assume that MAR conditions were met and MI was reliably employed.

Another factor that might have influenced our MI results is the number of imputed data sets used. Since estimating the values of a missing observation only once may be inaccurate, MI adopts multiple estimations to account for the uncertainty of each imputed value. However, discussion remains on how many estimates are necessary for obtaining accurate regression estimates. Although five imputation sets were reported to be sufficient for data sets with up
to 50% of missing values\textsuperscript{10-12} in some cases a much larger number of imputation data sets, i.e., 20 or more, might be necessary.\textsuperscript{13} In our study we ultimately generated 5 imputation datasets, however we examined the effect of applying more imputation sets. Since no significant changes were observed in regression coefficient estimates while using up to 25 imputation sets, we reported our results based on the previously recommended sufficient number of 5 imputation data sets.\textsuperscript{10-12} After applying a MI routine, a researcher is faced with multiple datasets that have to be combined in order to provide a single result. How these multiple data are handled might be of influence on the derivation and composition of the prognostic model. In each of the imputed data sets a prognostic model can be fitted and a variable selection method, such as backward selection, can be employed. However, resulting models can differ regarding selected predictors and regression coefficient estimates. How predictor selection should be performed with multiple imputed data is not yet completely clear.\textsuperscript{14} Rubin's rules are still regarded as the best alternative.\textsuperscript{14,15} However, this approach is very computationally intensive since each model selection step involves fitting the model over all imputed data sets and combining estimates across these models. Wood et al.\textsuperscript{14} described and tested a computationally simple two step procedure which they found to be “most useful”. We therefore adopted a similar two-step procedure. First the most occurring predictors (≥40%) from the 5 imputation sets were selected. In the second step, the contribution of these predictors to a pooled model was tested by a likelihood ratio test.\textsuperscript{16} In this manner we derived a single prognostic model from our 5 imputed data sets.

**Model stability in prognostic research**

In all models derived by regression analysis (\textit{chapters 2 and 3}) we employed a predictor selection procedure based on statistical
significance. This means that from a large group of potential predictors a smaller subset of predictors was selected by using univariable and multivariable, i.e., backward selection, resulting in a final optimal model. Although commonly employed in prognostic research, this methodology is often discussed, criticized or even advised against\textsuperscript{17} since resulting models might not reflect the “optimal” model. The reasons for this are: Statistical predictor selection is known to have difficulties in including only the most important predictors. The use of a threshold value for predictor inclusion or retention raises the chance of encountering a type I error, meaning that variables that are less important and do not contribute much to the predictive value and performance of the model, will be included on basis of their overoptimistic high value and statistical significance level. This process of variable selection leads to models that are data dependent so that the model selected might depend on a small number of observations in the data. It may lead to the identification of highly unstable models, models that have a different number and type of variables when the process of variable selection is repeated in other slightly different patients. The ‘best’ way of selecting potential predictors to be included in a multivariable model, especially in situations where consistent subject matter knowledge is lacking as was the case in our study, is still unresolved.\textsuperscript{18}

In chapter 3 we showed how prognostic modelling might benefit from the additional assessment of the stability of a multivariable model. For this we applied a bootstrap model selection procedure.\textsuperscript{19,20} In this procedure the bootstrap is expected to mimic sampling variation with the idea that the population from which the derivation sample was drawn is one of the many possible samples that can be used to build a model. Fitting a model in the bootstrapped data is expected to select variables taking sampling
general discussion

variation in to account,\textsuperscript{19,21,22} thus mimicking that the model is built in several comparable groups of patients which consult the GP with new episodes of shoulder complaints. Repeating this procedure a large number of times (500 times) provides information on which group of predictors is consistently included in the selected models and helps identifying the most informative model. This procedure can also be applied in combination with Multiple Imputation, as we did in chapter 3. However, which criteria should be used in order to evaluate model stability and when a model can be assumed stable is not yet clear.

**Classification and Regression Trees as alternative modelling method**

Besides the commonly employed logistic regression analysis, alternative methods have been described in the literature for deriving prediction models. One commonly used alternative technique is the Classification And Regression Tree (CART) method, which is a form of recursive partitioning where the derivation data is subsequently split into smaller subgroups so that a tree like diagram is created.\textsuperscript{23} According to proponents of this methodology, this tree like structure makes for more clinically intuitive models which therefore are better suited for application in clinical practice than regression based models.\textsuperscript{24}

When comparing both methods for the derivation of a model to predict persistent shoulder pain (chapter 4), we showed that a regression based model was more (internally) valid than a model derived by the CART technique. These results were based on applying a bootstrap internal validation procedure which has become standard in regression based prognostic research, but to our knowledge was never before used to estimate overoptimism for
a CART model. The largely analogous method of cross-validation, which is more commonly applied in recursive partitioning, however has been described to give comparable results regarding the superior internal validity of regression models over CART models.\textsuperscript{25} This implies that in comparison to regression based models, CART models can indeed be more dependent on the characteristics of the derivation data and are therefore less suited for application in new patients.\textsuperscript{26} Therefore, we chose not to externally validate our derived CART model.

**Clinical determinants of persistent shoulder pain**

In our Delphi study (\textit{chapter 5}), clinicians mentioned the following determinants as key predictors for persistent shoulder pain: longer symptom duration, higher pain catastrophizing, presence of widespread symptoms, more fear-avoidance beliefs, coexisting neck pain, higher severity of shoulder disability at baseline, multisite pain, higher age, higher pain intensity and more illness perceptions. The role of psychosocial determinants (pain catastrophizing, fear-avoidance beliefs and illness perceptions) has not been reported extensively in the shoulder pain literature, and evidence on the prognostic value of these determinants for shoulder complaints is limited.\textsuperscript{27,28} Our clinical consensus study thus provided indications on the importance of psychosocial factors as possible determinants for the prognosis of persistent shoulder complaints.

All predictors, including the psychosocial factors, that were put forward by clinical consensus closely resembled factors that were previously identified as potential prognostic factors for the assessment of any regional musculoskeletal pain complaint.\textsuperscript{29} It is known from the literature that the apparently different regional musculoskeletal pain complaints share a range of generic factors.\textsuperscript{29} Although our expert panel consisted primarily of health care
professionals with expertise on shoulder pain, it is most likely that they also have clinical experience with other regional musculoskeletal complaints. This expertise with other regional musculoskeletal complaints might be reflected in our consensus result by the inclusion of the psychosocial determinants pain catastrophizing, fear-avoidance beliefs and illness perceptions as key predictors for persistent shoulder pain.

Before discussing the impact of these findings for clinical practice, it is important to consider the method for deriving clinical consensus. Obtaining consensus from a group of clinical experts can be a challenging process. Individual opinions might diverge strongly. To unite multiple opinions into a commonly shared result, i.e., consensus, a structured communication within a group of experts is required. In our research, we applied such a structured approach by using a Delphi consensus method. By using subsequent web-based questionnaires, we consulted a group of experts involved in or having thorough knowledge of the assessment and management of (patients with) shoulder complaints regarding the key predictors for persistent shoulder pain. After each questionnaire we fed the results back so that group members could reflect upon them and individual opinions could converge to a mutual shared result. In this process, bias may have occurred since a Delphi procedure is susceptible to three forms of bias; 1) selection bias, 2) subject bias and 3) bias caused by interpretation of the results by the researcher.

Since a Delphi procedure has no geographical or size limits, our expert panel could be large (n=41), consisting of an international (United Kingdom and the Netherlands) and multidisciplinary group of experts (GPs, physiotherapists, rheumatologists, epidemiologists, a manual therapist and a senior lecturer in occupational medicine). Because of this heterogeneity in our expert panel, the occurrence of selection bias, i.e., lack of broad input and discussion due to shared
Opinions was highly unlikely. In addition, the probability of achieving high quality and widely acceptable results increases when supported by a heterogeneous panel.

During the subsequent rounds of our Delphi procedure, panel members adjusted their opinion so that individual views converged to a shared result. This phenomenon is characteristic for a Delphi study, however it could be interpreted as subject bias. Subject bias occurs when peer pressure is the main reason for adjusting one’s view, and is less likely when individual opinions are adjusted based on strong convincing arguments. Panel members did not know each other’s identity, so every individual contribution was equally important and no dominant or authoritarian opinions, psychological factors, vindictive disputes or diverse levels of knowledge could affect the results. This anonymity in our Delphi procedure helped to prevent bias.

One of the largest objections concerning the Delphi procedure is that a researcher can introduce bias by interpreting the results. In our Delphi procedure we tried to prevent this bias by a priori formulating consensus criteria, keeping the procedure as transparent as possible, feeding back all reasoned panel opinions, elucidating and presenting all of our interpretations for consensus, and making sure that the final say in all decisions was always with the panel.

All in all, we are confident that our expert panel technique was appropriate for obtaining reliable results, and that major sources of bias did not hamper the Delphi procedure itself. Our results are therefore likely to reflect the current clinical knowledge of health care professionals on key determinants for persistent shoulder complaints.
Clinical prognosis of persistent shoulder pain

In *chapter 6* we compared the expert-based and statistical model to the GPs’ own estimation of the likely outcome of the shoulder pain problem. The results were notable in several ways: the GPs underestimated the risk of persistent shoulder pain after 3 months. The statistical and expert-based prognostic model performed slightly better. However, the performance of both prognostic models in a new cohort of patients with a new episode of shoulder complaints was also quite disappointing. Of interest was that the expert-based model performed better than the statistical model according to their discriminative abilities.

The fact that GPs’ estimates of the likely outcome performed less well compared to the expert-based prognostic model may seem surprising given the fact that the Delphi procedure was designed to derive a model that would more closely reflect the implicit process of clinical reasoning, and would incorporate prognostic information considered important by clinicians. There may be several explanations for this finding.

Firstly, the importance of the predictive factors might differ for each patient. In a clinical prediction model this cannot be taken into account.

Secondly, the way in which we measured the predictive factors might be different from the way GPs incorporate information in their prognosis. As a last step in our Delphi procedure, we presented proposals for measuring the predictors and suggested cut-off points. The experts all readily accepted these. However, our panel included GPs and therapists with a specific interest in shoulder pain, and may not fully represent the GPs participating in our validation study. Furthermore, assessment of prognostic information in routine clinical practice is likely to be different from
using a standardised set of questions.

And thirdly, the process of integrating information in a GP’s mind cannot be reproduced by a logistic regression analysis.

Most of these reasons would be detrimental to the performance of the expert-based model, however, the expert-based model still performed better than the GP estimate. We believe this might be partly explained by our multidisciplinary group of experts enrolled in our Delphi study, and the brief GP consultation in routine primary care. Knowledge incorporated in the expert-based model is likely to be more elaborate and results therefore did not reflect the opinion of a GP estimating outcome based on a 10 minute consultation. Next, the clinical feasibility of our outcome definition “improvement of less than 50% in shoulder pain intensity score” might be low. Instead of predicting decreasing pain scores it might be more relevant for GPs to estimate more apparent outcomes such as expected overall recovery or the actual level of pain at follow-up.

It turned out that the GPs were too optimistic about the prognosis of their patients. While only 13% of all patients received a GP predicted probability higher than 50%, in reality 45% of all patients reported having persistent shoulder pain after three months. We think that GPs estimated the patient’s prognosis by also taking the potential effect of the prescribed treatment into account. Due to a very optimistic attitude towards treatment outcome, GPs possibly underestimated the patients’ probability of persistent shoulder complaints. Such a highly optimistic attitude towards treatment effects is not directly instigated by the Dutch GP guidelines on shoulder complaints, which indicate on the high rates of recurrent or persistent complaints among patients with shoulder complaints independent of treatment. The optimistic GP estimates might be the result of a positive attitude towards the patients and expressing hope for recovery. However, it is also possible that GPs really are not able to predict whether their
treatment will work or not, and what the likely outcome of the shoulder problem is.

Whatever the reason for the poor prediction by the GPs, it implies that the GPs might probably benefit from the use of clinical prediction rules. However, the models’ predictive ability in new patients foremost needs to be of sufficient magnitude, which in our developed models is not the case.

**Main conclusions**

- Dealing with missing data by Complete Case Analysis instead of the preferred Multiple Imputation can lead to considerable bias as illustrated by analysis of data from an empirical study.
- A bootstrap model selection procedure is an informative addition to the derivation of a prognostic model since it provides information on the stability of the selected model.
- The alternative modelling method of Classification And Regression Trees (CART) was less suited for the development of a prognostic model predicting persistent shoulder pain than regression analysis, since resulting models were far more optimistically estimated than regression based models.
- By conducting a Delphi study amongst professionals involved in the management of shoulder complaints ten key predictors for persistent shoulder pain were identified, from these ten, three (psychosocial) key predictors (pain catastrophizing, fear-avoidance beliefs and illness perceptions) were revealed for which evidence on their prognostic value for shoulder pain in the medical literature is limited.
- The transformation of clinically identified key predictors to an expert-based prognostic model led to a model that in a new cohort of patients with shoulder complaints showed better
predictive abilities compared to a statistically multiple imputation based model

- Predicting persistent shoulder pain at 3 months after initial consultation is difficult: Dutch general practitioners were not very successful in predicting persistent shoulder pain and our derived statistical and expert-based prognostic models performed poorly regarding the discrimination between patients with and without persistent shoulder pain

- External validation is required by assessing model performance in a new cohort of patients, even when they meet similar selection criteria as the derivation cohort, as models may show a considerable drop in performance from estimates derived by internal validation

**Recommendations for research**

- When deriving a prediction model in the presence of missing data use Multiple Imputation to minimize the loss of information due to missing values

- When performing Multiple Imputation the consequences of decisions regarding the number of imputation files, the method of predictor selection, and how results from multiple imputed data sets will be combined need to be addressed and reported

- Always examine model stability to assess and provide information on the problems caused by predictor selection

- When deriving a CART model always assess internal validation by cross-validation or bootstrapping after the model has been pruned for overoptimism since resulting models might still be very optimistically estimated

- When deriving a prediction model internal validation is required to estimate a model’s overoptimism. However as internal
validation is not per se indicative for performance in new shoulder pain patients, external validation is always needed

Since in the literature the number of reported musculoskeletal prediction increases, future prognostic research should focus on how results can be synthesized in order to obtain a consistent and informative prognostic factors and investigate whether these could also be applied to predict persistent shoulder complaints
References


36. Winters JC, Van der Windt DAWM, Spinnewijn WEM, De Jongh AC, Van der Heijden GJMG, Buis
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 straightforward procedure, in the process choices need to be made and the effects of these choices need to be evaluated and reported.
SAMENVATTING

In de eerstelijnsgeneeskunde en de huisartsgeneeskunde in het bijzonder, worden clinici vaak geconfronteerd met patiënten met bewegingsapparaat gerelateerde klachten. De behandeling van deze klachten blijkt echter vaak lastig, klachten houden lang aan of zijn wederkerig en duidelijke diagnostische criteria ontbreken. Om clinici te helpen bij de inschatting van de ernst van de klachten heeft de wetenschap zich de laatste tijd gericht op predictie onderzoek. Hierin wordt getracht om factoren te identificeren welke een ongunstige uitkomst of beloop van de aanwezige klachten kunnen voorspellen. Deze factoren kunnen vervolgens gecombineerd worden tot een rekenregel ofwel een predictie model. Echter, het vinden van een eenvoudig predictie model dat voor alle patiënten consistente voorspellingen geeft blijkt erg lastig. Enkele methodologische kwesties zoals missnede waarden en model stabiliteit bemoeilijken het vinden van een dergelijk model. Met het in dit proefschrift beschreven onderzoek proberen we inzicht in deze verstorende factoren te krijgen en proberen we uit te zoeken hoe we hun invloed kunnen minimaliseren. Hiertoe zullen verschillende analyse methoden in verschillende bewegingsapparaat gerelateerde data sets toegepast worden om zo meer kennis te vergaren over de optimale methode om tot een valide predictie model te komen.
In *hoofdstuk 1* beschrijven we de doelstellingen en het kader van dit proefschrift. Daarnaast bespreken we de epidemiologie van bewegingsapparaten gerelateerde klachten in het algemeen, geven we aan hoe predictie regels van toegevoegde waarde kunnen zijn bij de behandeling van bewegingsapparaten gerelateerde klachten in de eerstelijnsgeneeskunde en benadrukken we hoe methodologische kwesties zoals missende waarden en model stabiliteit de ontwikkeling van klinisch bruikbare predictie regels in de weg kan staan.

In *hoofdstuk 2* bieden we onderzoekers een empirische illustratie van het omgaan met missende waarden door deze te negeren en zodoende alle respondenten met missende waarden uit de data te laten (Complete Case Analyse, CCA) of door de waarde van de missende informatie te schatten (Multiple Imputatie, MI). Hiertoe gebruikten we data van het BeBack (Beliefs about Backpain) cohort, een studie naar de psychologische invloeden op het aanhouden van rugpijn onder 1591 eerstelijns patiënten met rugpijn in het Verenigd Koninkrijk. Deelnemers dienden voor elk meetpunt (voor- en nameting) apart toestemming te verlenen. Naast missende waarden in de voormeting (14%) leverde dit ook een behoorlijke uitval (51%) ten tijde van de nameting op. Wanneer de deelnemers met incomplete data vergeleken werden met de deelnemers waarvan wel alle informatie bekend was, viel het op dat er data selectief ontbrak. Deelnemers met complete data verschilden wat betreft de verdeling van voorspellers en uitkomst ten opzichte van deelnemers met incomplete data. Het gevolg hiervan was dat de predictie modellen ontwikkeld met alleen complete data of alle data (met invulling van de missende waarden) van elkaar verschilden wat betreft model samenstelling, voorspellend vermogen en overoptimisme. Hiermee zijn onze resultaten illustratief voor hoe CCA in het geval van missende waarden de ontwikkeling van een predictie model
en hoe MI deze verstoring minimaliseert. Daarom adviseren we in het geval van missende waarden multiple imputatie te gebruiken bij het ontwikkelen van een predictie model.

Bij het gebruik van multiple imputatie moet echter wel rekening gehouden worden met twee verstorende factoren: 1) het tot één model combineren van de meerdere geïmputeerde data sets en 2) de extra instabiliteit geïntroduceerd door het herhaald schatten van de missende waarden. Met het oog op deze verstorende factoren onderzochten we in hoofdstuk 3 of een bootstrap model selectie procedure een waardevolle toevoeging aan multiple imputatie is. Hiertoe werd deze techniek toegepast in een data set afkomstig uit het Nederlands Schouder Onderzoek (NSO), een cohort bestaand uit 587 schouderpijn patiënten afkomstig uit de Nederlandse huisartsenpraktijk. Door in afzonderlijke stappen de sterke van de zwakke voorspellers te scheiden en vervolgens in meerdere bootstrap samples alle mogelijke combinaties van voorspellers te onderzoeken verkregen we een beeld van de robuustheid van het geselecteerde predictie model.

In hoofdstuk 4 vergelijken we een van de meest gebruikte alternatieve technieken om tot een predictie model te komen: Classification And Segression Tree (CART) analyse, met de doorgaans gebruikte logistische regressie methode. Om te bepalen welke van de twee beter geschikt is om een model te ontwikkelen pasten we beide technieken toe in de data van het Nederlands Schouder Onderzoek (NSO), een cohort bestaand uit 587 schouderpijn patiënten afkomstig uit de Nederlandse huisartsenpraktijk. We vergeleken de resultaten van beide methoden op verschillende belangrijke punten in de ontwikkeling van een predictie model. Het totale aantal geïncludeerde voorspellers was voor beide modellen gelijk, echter de exacte model
methodologische kwesties van klinische predictie regels voor schouderpijn in de huisartsenpraktijk

samenstelling verschilde onderling. Het voorspellend vermogen in de derivatie data was vrijwel gelijk voor beide modellen (gelijke $R^2_N$ van 19% en een AUC van 0.72 voor logistische regressie en 0.70 voor CART). Echter, bij toepassing in vergelijkbare patiënten (interne validatie) bleek het CART model slechter te presteren dan het logistisch regressie model. Op basis van deze bevindingen concluderen we dat ons logistisch regressie model beter geschikt is voor het voorspellen van aanhoudende schouderklachten dan het door ons ontwikkeld CART model.

In hoofdstuk 5 trachtten we middels een groep van experts op het gebied van de behandeling van schouderklachten tot consensus te komen wat betreft de belangrijkste voorspellers van aanhoudende schouderpijn. Vervolgens testte we het voorspellend vermogen van deze klinische voorspellers. We stelden een internationaal (Verenigd Koninkrijk en Nederland) en multidisciplinair panel samen van 41 experts op het gebied van schouderklachten. In drie opeenvolgende Delphi ronden selecteerde het expert panel de volgende voorspellers voor aanhoudende schouderpijn: klachtenduur, catastroferen, klachten in het verleden, angst vermijding, bijkomende nekpijn, de ernst van bewegingshinder in de schouder, pijn op meerdere lichaamspuncties, leeftijd, de intensiteit van de schouderpijn en ziekteperceptie. Met deze voorspellers werden er, met gebruik van de NSO data, twee klinisch gebaseerde predictie modellen ontworpen. Eén waarin de informatie zoveel mogelijk als continue voorspellers werd opgenomen en één waarin het grootste gedeelte van de informatie als dichotome voorspellers werd opgenomen. Wanneer vergeleken met een eerdere ontwikkeld statistisch model (hoofdstuk 3) bleek de prestatie van de klinisch gebaseerde modellen iets slechter (statistisch model AUC=0.702, klinisch dichotoom model AUC=0.656 en klinisch continu model AUC=0.679). Al verschilden de klinische modellen van het statistische model wat
samenvatting

betreft model samenstelling, we concluderen dat beide modellen bevestigen dat klachtenduur, de ernst van bewegingsbeperkingen van de schouder tijdens de voormeting en de intensiteit van de schouderpijn tijdens de voormeting belangrijke voorspellers voor aanhoudende schouderpijn zijn. Daarnaast concluderen we dat de verschillen in interne validiteit tussen de klinisch gebaseerde modellen en het statistische model bevestigd dienen te worden door beide modellen toe te passen op nieuwe schouderpatiënten.

In hoofdstuk 6 onderzochten we de externe validiteit van de twee meest belovende door ons ontwikkelde modellen, het klinisch gebaseerde continue model uit hoofdstuk 5 en het statistisch ontwikkelde MI-5 model uit hoofdstuk 3. Daarnaast vergeleken we het voorspellend vermogens van beide modellen met de prognose van de huisarts zelf welke gesteld werd tijdens het consult. In 23 huisartspraktijken rekruteerden we 203 patiënten met een nieuwe episode van schouderpijn. Deze patiënten vormden het SchouderPijn Prognose (SPP) cohort. In dit cohort bleken het statistische en klinisch gebaseerde model beiden suboptimaal te voorspellen met AUC waarden van respectievelijk 0.580 en 0.625. De inschatting van de toekomstige staat van schouderpijn van de 35 huisartsen in het onderzoek bleek met een AUC van 0.506 echter nog lager. Hiermee geven onze resultaten aan hoe moeilijk het is om het aanhouden van schouderklachten na drie maanden bij patiënten in de huisartsenpraktijk te voorspellen. Onze resultaten suggereren dat predictie modellen wellicht van toegevoegde waarde zouden kunnen zijn in de huisartsenpraktijk, echter de door ons ontwikkelde modellen bleken niet valide genoeg om een echte toevoeging te bieden.

In hoofdstuk 7 plaatsen we de resultaten van dit onderzoek in een breeder perspectief, sommen we de voornaamste conclusies van dit
proefschrift op en doen we aanbevelingen voor predictief onderzoek. In grote lijnen zouden onderzoekers op het gebied van predictie onderzoek altijd op de hoogte moeten zijn van de verstorende invloed van missende waarden en model stabiliteit. De ontwikkeling van een predictie model is geen recht toe recht aan exercitie, in het proces dienen belangrijke keuzes gemaakt te worden en het effect van deze keuzes op de resultaten dienen geëvalueerd en gerapporteerd te worden.
DANKWOORD
WORD OF THANKS
DANKWOORD / WORD OF THANKS

In de regel wordt een proefschrift van achter naar voren gelezen. Waar u hoogst waarschijnlijk begint met het lezen van dit dankwoord, komt voor mij met het schrijven van dit dankwoord mijn promotie periode tot zijn einde. Een bewogen periode welke zonder de steun, begeleiding en medewerking van velen niet tot dit resultaat zou hebben geleid. Daarom wil ik bij deze mijn dank betuigen aan iedereen die de totstandkoming van dit proefschrift mede mogelijk heeft gemaakt.
Op de eerste plaats wil ik graag mijn promotoren en co-promotoren bedanken voor al het werk dat zij verricht hebben en zo voor een groot deel verantwoordelijk zijn voor het slagen van dit project.

Daniëlle, bedankt voor je vertrouwen om mij als groentje aan dit project te laten beginnen. Vanuit jouw enorme kennis en kunde heb je mij op een hele prettige en informele wijze wegwijsgemaakt in de wereld van de epidemiologie, predictie en van de schouderklachten. De unieke overleggen vanuit de deurposten van onze tegenoverliggende kamers zal ik niet snel vergeten.

Riekie, na de inauguratie van Daniëlle in Keele nam jij het stokje over en je werd de motor in de voortgang van het project. Hiervoor ben ik je heel erg dankbaar net als voor de aanmoedigingen wanneer ik het even niet meer zag zitten, inzichtelijke toevoegingen aan de manuscripten en het altijd lezen van de stukken zelfs wanneer je ze op het aller laatste moment had ontvangen.

Martijn, als promotor van het eerste uur wil ik je bedanken voor jouw onvoorwaardelijke inzet en de tijd die je toch steeds weer in je drukke agenda wist te vinden. Daarnaast had ik me geen betere gids door het labyrint dat methodologie van predictieregels heet kunnen wensen.

Henriëtte, jouw klinische blik en expertise als huisarts en onderzoeker was van onschatbare waarde voor dit project. Tevens was je een prima promotor met oog voor het grote plaatje alsmede voor de kleine details en voor mijn persoonlijke ontwikkeling als promovendus. Bedankt.

Omdat je zonder data geen onderzoek kunt doen wil ik graag iedereen bedanken die bijgedragen heeft aan de gegevensverzameling van de SchouderPijn Prognose Studie.

Als eerste natuurlijk alle Personen uit het SPP cohort heel erg bedankt voor uw deelname aan dit onderzoek waarbij we u enkele malen het hemd van het lijf mochten vragen enkel in ruil voor een simpele balpen.

Daarnaast ben ik de deelnemende huisartsen aan de SPPS heel erg dankbaar voor de tijd die ze voor mij maakten, het informeren en rekrutering van potentiële deelnemers en voor het aanleveren van de prognose schattingen.

Bij de benadering van het Gros van deze huisartsen was de hulp van Petra Elders en Valentina Blom van het Academisch Netwerk Huisartsgeneeskunde onontbeerlijk. Dank voor jullie inzet.

Marijke van Dijk en Marlieke van der Eerden, heel erg bedankt voor de ontelbare telefoontjes naar huisartsen en patiënten, de vele verstuurde brieven, het bijhouden van de patiënten administratie en nog veel meer klusjes waarmee jullie mij wat werk uit handen namen.

Daarnaast wil ik graag alle panelleden uit de Delphi studie van harte bedanken voor hun verrijkende en leerzame input. Since this Delphi study was an international collaboration of experts in the field of shoulder disorders, I also would like to thank all panel members in English for their contributions and insightful comments.

Other international contributors I would like to thank are Nadine Foster, Kate Dunn, George Peat and Peter Croft.

Nadine and Kate, thank you for putting the data from the BeBack cohort at my disposal and for their contributions to the manuscript that led to the second chapter of this thesis. Your critical appraisal really improved the quality of this work.
George and Peter, many thanks for your contributions to the third chapter of this thesis. I found your intelligible observations and remarks truly motivational.

Ton Kuijpers, heel veel dank voor het gebruik mogen maken van jouw NSO cohort en voor het kritisch beoordelen van de manuscripten welke uiteindelijk de afzonderlijke hoofdstukken van dit proefschrift zouden vormen.

Alle collega’s van de afdeling huisartsgeneeskunde, epidemiologie en biostatistiek en de predictie werkgroep bedankt voor de fijne werksfeer.

Speciale dank gaat uit naar oud kamergenoten Marcel, Jeroen en Uriëll voor de broodnodige afleiding. Al zat ik vaak achter mijn pc gekluisterd, jullie kregen het toch voor elkaar om mij af en toe van het toetsenbord los te weken en even de gedachten op iets anders te brengen waarna er weer fris verder gewerkt kon worden.

Tobias, onze gesprekken over predictie waren altijd erg verrij kend en zo werd het beoordelen van artikelen voor jouw review een prettige ervaring.

En als laatste maar zeker niet als minste wil ik mijn ouders en Suzanne bedanken voor hun steun in de afgelopen jaren. Ik weet zeker dat jullie minstens net zo blij en trots zijn als ik nu dit boekje af is.

Lieve Suus bedankt dat ik altijd bij je terecht kon wanneer het niet liep zoals ik zou willen. Nu je dit leest is het werk echt echt af! Dank je voor je wijze raad, liefde, humor en steun.
```
B <- function(data, outcome, name, m, alpha) {
  # functions
  logistic <- function(mu) exp(mu)/(1-exp(mu))
  boot1 <- function(x, N, formula, Data, alpha) {
    modpars <- 0
    for (i in 1:m) {
      index.i <- sample(x, N, replace=T)
      bdata.i <- Data[index.i[1:N], ]
      fct.i <- t.test (formula, data=bdata.i)
      bmodpars[i, ] <- fct.i[2, Nmax]$t
      bmodpars <- c(bmodpars, fct.i[2, Nmax]$t)
    }
    pm <- matrix(0, a, a)
    for (k in 1:a) {
      for (j in 1:length(bmodpars[[k]])) {
        pm[k, j] <- bmodpars[[k]][j]]
      }
    }
    count.row <- function(x)
      order.x <- order(as.data.frame(x))
      equal.count <- equal.count(x, order=1, j=1)
      count <- length(unique, rownames(count.row<which(x, logical=T, equal=T, previous)), tf)
      rownames(count) <- which(x, logical=T, equal=T, previous)
      count <- count[[order(unique, rownames(count)])]
    # predictor selection
    Data <- read.csv((data, use.value.labels=T, na.string="NA", row.names="NA", sep="", quote=T, collapse="|")
    x <- as.numeric(row.names (Data))
    N <- length(formula <- parse(paste((outcome, "~", paste(x, collapse="|"))
    pm <- boot1(x, N, formula, Data, 1000)
    paramfreq <- 0
    for (i in 1:i) {
      paramfreq[1] <- sum(pm[[1]]
    }
    # model selection
    names <- names(c(wha, names <- names, formula <- as.formula(N <- length(names) <- as.numeric(row.names <- length(names)
    count.row <- count.row <- which(x, formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(formula <- as.formula(fit1 <- lm(formula, data=Dat
fit2 <- glm(formula, data=Dat
coef(fit1)
preblue <- logistic(x, c1, x, c1, pmodpars, (fit1)[c1]
return(list (predictor=1)
```