Statistics
in Medicine

Research Article

Received 10 May 2015, Accepted 13 June 2016 Published online 11 July 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7034

Semi-varying coefficient multinomial logistic regression for disease progression risk prediction

Yuan Ke,[a] Bo Fu[b,c*†] and Wenyang Zhang[d]

This paper proposes a risk prediction model using semi-varying coefficient multinomial logistic regression. We use a penalized local likelihood method to do the model selection and estimate both functional and constant coefficients in the selected model. The model can be used to improve predictive modelling when non-linear interactions between predictors are present. We conduct a simulation study to assess our method's performance, and the results show that the model selection procedure works well with small average numbers of wrong-selection or missing-selection. We illustrate the use of our method by applying it to classify the patients with early rheumatoid arthritis at baseline into different risk groups in future disease progression. We use a leave-one-out cross-validation method to assess its correct prediction rate and propose a recalibration framework to evaluate how reliable are the predicted risks. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: model selection; multinomial logistic regression; penalized likelihood; risk prediction; varying coefficients

1. Introduction

The motivation for this paper arose from a medical study, where the research interest is to classify the patients with early inflammatory polyarthritis to three groups at different risk of progression to functional disability by using their baseline information. The outcome of interest is a patient's functional disability status at the end of a 5-year follow-up, which is a discrete variable having three levels (low risk, moderate risk and high risk) of disability. Such a risk prediction model is important in medical application in order to identify a subgroup of patients at early stage of disease onset who are at higher risk to progress to a worse outcome so that more aggressive treatment strategies, such as use of biologic therapy, could be matched to them [1–3].

Logistic regression models are widely used for developing predictive models where the outcome of interest is a dichotomous or nominal-scaled variable. When the outcome variable can take more than two values, a multinomial logistic regression is usually applied [4]. New multi-category classification methods in multinomial logistic regression were recently discussed in Li *et al.* [5].

However, usual logistic regression models assume that the effects of predictors on outcomes are constant. We relax this assumption by incorporating a varying coefficient structure to allow the effects of the predictors to vary smoothly with the change in a single-continuous covariate U, such as baseline disease duration in our motivating example. Varying-coefficient models were introduced by Cleveland *et al.* [6] as a useful nonparametric tool to analyze complex dynamic data [7–14]. In particular, the use of such a nonparametric structure permits non-linear interactions between the predictors and a particular variable,

[a] *Department of Operational Research and Financial Engineering, Princeton University, Princeton, NJ 08540, U.S.A.*
[b] *Administrative Data Research Centre for England and Institute of Child Health, University College London, London NW1 2DA, U.K.*
[c] *Centre for Biostatistics and Arthritis Research UK Epidemiology Unit, The University of Manchester, Manchester M13 9PL, U.K.*
[d] *Department of Mathematics, The University of York, York YO10 5DD, U.K.*
[*] *Correspondence to: Bo Fu, Administrative Data Research Centre for England, Farr Institute of Health Informatics Research, University College London, 222 Euston Road, London NW1 2DA, U.K.*
[†] *E-mail: b.fu@ucl.ac.uk*

which could be useful to improve the model fitting [7, 15, 16]. For example, in Section 4, we will consider a relatively large number of candidate covariates at baseline for predicting future progression to functional disability in patients with early inflammatory polyarthritis. The set of candidate covariates may include patients' demographic factors (e.g. age and gender), serological and genetic factors (e.g. rheumatic factor status and number of copies of shared epitope), disease activity and severity measures (e.g. number of swollen or tender joints), social-economic factors (e.g. index of multiple deprivation score) and so on [3]. The number of potential predictors could be large if more biomarkers are available at baseline. It is of interest to allow the effects of some baseline predictors to depend on the disease duration from disease symptom onset to the baseline time when predictor variables were measured. By incorporating flexible interactions between the effects of baseline predictors and disease duration in the prediction model, we account for influences due to variations over time from disease onset to baseline between subjects. The research questions are then (i) which variables among a large number of candidates should be selected in the predictive model; and (ii) which variables have varying effects among the selected predictors.

Variable selection is an essential part of statistical analysis to improve model predictability. Failing to select predictors which are associated with an outcome will lead to bias. On the other hand, the prediction power may be reduced if we include variables unrelated to the outcome. The coefficient estimates in a logistic regression could be biased if we include too many unrelated covariates. Traditional methods, such as stepwise deletion and best-subset selection based on AIC or BIC, are commonly used, but they tend to select an overfitting model in a logistic regression, leading to inaccurate scientific conclusion [17]. Also, traditional methods will be impractical if a large number of candidate predictors are included. They are computationally intensive and might be unstable because of their inherited discreteness [18].

When the number of candidate covariates is large, the traditional statistical methods can easily fail because of the so called 'curse of dimensionality' [19]. The new generation of variable selection methods based on penalized functions, such as ridge regression [20], the least absolute shrinkage and selection operator (LASSO) [21], smoothly clipped absolute deviation (SCAD) [22], group LASSO [23], adaptive LASSO [24], Minimax Concave Penalty [25], as well as Bayesian hierarchical models [26], have become particularly attractive in the analysis of high-dimensional data. These penalized methods can do the model selection and the parameter estimation simultaneously through a constrained optimization. The likelihood function and the constraints (usually expressed as penalties) together make a trade-off between the goodness of fit and the model complexity. Given an appropriate penalty function, the penalized methods can remove the unrelated covariates and reduce the dimensionality.

The modern variable selection methods based on penalized functions have also been extended to nonparametric models such as varying coefficient models. For example, Wang and Xia [11] proposed a K-LASSO method based on group penalization and quadratic approximation to handle the model selection for varying coefficient models. Lian [27] considered the variable selection for high-dimensional generalized varying coefficient models. Li *et al.* [28] proposed a variable selection and structure identification method for generalized semi-varying coefficient models. Other recent related work on variable selection includes Kong and Xia [29] for a single-index model, Tao and Xia [30] for an adaptive semi-varying coefficient model, Kuk *et al.* [31] for predictive modelling based on logistic regression and Stefanski *et al.* [32] for nonparametric classification methods.

In this paper, we consider variable selection for semi-varying coefficient multinomial logistic regression and introduce a multi-category grouping method for risk prediction in chronic disease progression. An attractive feature of our method is that it not only selects predictors but also select their coefficient types (constant or functional). It also allows the number of potential covariates to increase with sample size and is able to handle the case where the number of potential variables is large. This would be particularly useful in practice as we may include all available potential predictors to improve prediction accuracy. Our method contains two steps:

(1) **Model selection and parameter estimation**. We start with a full multinomial logistic regression model including all candidate covariates and apply a penalized likelihood approach to select predictors and the types of their coefficients (constant or functional). We then estimate both constant and functional coefficients for the selected model.
(2) **Prediction**. For each subject, we calculate the conditional probabilities that this subject belongs to different risk groups based on the selected model and its estimated coefficients. A subject is predictable if the maximum of the estimated group-membership probabilities exceeds a given threshold and is then classified to the corresponding risk group with the largest group-membership probability.

In the aforementioned procedures, Step 2 just involves some trivial calculations, and the key step for our modelling is the model selection part.

The rest of this paper is arranged as follows. In Section 2, we introduce the scientific motivation of this paper. In Section 3, we discuss the proposed risk prediction model in detail, which is based on the ideas of kernel smoothing, penalized likelihood, local linear approximation and penalization on deviations. The selection of tuning parameters is presented in Section 3.3. Section 4 focuses on an application to inflammatory polyarthritis data. In the end, Section 5 gives a brief discussion.

2. Motivating data example

2.1. *Scientific motivation*

Our work was motivated by an analysis of a medical dataset from a primary care-based prospective cohort of patients with recent onset inflammatory polyarthritis [33]. Rheumatoid arthritis (RA) is the most common inflammatory disease of the joints, which is associated with progressive joint destruction resulting in severe disability. However, it is difficult to identify RA at an early stage of disease onset because no tests or diagnostic criteria are available to define early RA [34]. A lab test that often helps with diagnosis of RA at a follow-up stage is anti-cyclic citrullinated peptide antibody test. Early arthritis may be progressed into established RA or another definite arthritis disease or may remain undifferentiated. To better manage the outcome in arthritis, it has been suggested by clinical researchers to first recognize inflammatory arthritis and then estimate the risk of developing persistent and erosive irreversible arthritis such as RA in order to propose an optimal treatment option [35].

Rheumatoid arthritis is a very heterogeneous disease in terms of disease progression outcome. Some patients with RA do not develop any severe outcome, such as erosion, even after a long time, but the majority will have bone erosion and cartilage breakdown resulting in joint destruction and functional disability. For the management of early inflammatory polyarthritis, the European League against rheumatism recommends that patients at risk of developing persistent and/or erosive arthritis should be started with disease-modifying anti-rheumatic drugs (DMARDs) as early as possible, even if they do not yet fulfil established classification criteria for RA [36]. Furthermore, the revolutionary introduction of biologic agents, such as anti-tumour necrosis factor, in the past decade offers patients a new and very effective treatment option alternative to the traditional DMARDs. Early treatment with biologic agents has been shown by published studies to improve clinical outcomes, patients' functional status and health-related quality of life [37]. However, biologic agents have potential to leave the patients more vulnerable to severe adverse events such as infection or malignancy because anti-tumour necrosis factor is involved in many aspects of host immunity [38]. Also, the drug costs of treatment with biologic agents are much higher compared with DMARDs.

In order to achieve the goal of personalized treatment and optimal early use of biologic agents in the management of RA, it is necessary to identify a subgroup of patients at baseline who are at higher risk to progress into a worse functional status in future or have better response to biologics treatment so that specific treatment strategies are matched to individual patients. In this paper, our scientific interest focuses on a prediction model to classify the patients into groups with different risk of progression to severe outcome rather than different responses to treatment. An ideal therapeutic strategy should then be based on such an appropriate prediction of the disease progression risk [36]. The aim of this study is to improve the predictive modelling by identifying significant prognostic factors (or predictors) associated with disease progression together with their significant interactions.

2.2. *Health Assessment Questionnaire progression data*

The data sample we study comprises 290 patients, who were recruited to the arthritis register cohort between 1990 and 1994 and have disease duration from symptom onset to registration less than 3 years. The disease outcome of interest is functional disability status, which is an important clinical measure in RA as it has been shown to be predictive of crucial RA-related outcomes, such as mortality. This measure was assessed using the modified British version of the Health Assessment Questionnaire (HAQ) score. The questionnaire comprises 20 questions in eight categories. Each question is given a score of 0 (no difficulty), 1 (some difficulty), 2 (much difficulty or need of assistance) or 3 (unable to perform). The score for each category is determined by the highest score in that category, and the sum of scores is then divided by the number of categories, yielding a total HAQ score ranging from 0 (best) to 3 (worst). All patients in our study sample have mild disease outcome at registration (baseline) with baseline HAQ scores between 0 and 1 and were followed for at least 5 years. The response variable Y is the functional disability status at the end of a 5-year follow-up since registration. $Y = 1$, if the functional disability status at the end of follow-up is at low risk (HAQ score between 0 and 1); 2 if the functional disability status is at moderate risk (HAQ score between 1 and 2); and 3 if the functional disability status is at high risk (HAQ score between 2 and 3). Such a classification into three groups based on HAQ score was suggested by Wolfe *et al*. [39], and it provides important and clinically useful current and predictive information regarding RA status, utilization of services and mortality [39].

In the predictive model, the candidate predictors include age at registration, gender, number of swollen joints out of 51 joints, number of tender joints of 51 joints, rheumatic factor (1= *positive* or 0= *negative*), smoking status (three categories: non-smoker, current smoker or ex-smoker), socio-economic status defined as an area-level category variable based on the nationally determined quartiles of the index of multiple deprivation score used in the UK (four categories: least deprived group, lower middle deprived group, upper middle deprived group and most deprived group) [3], number of copies of the shared epitope which is an established genetic biomarker in RA, fulfilment of the American College of Rheumatology 1987 classification criteria for RA (1 = *yes* or 0 = *no*), season of birth (four categories: spring, summer, autumn or winter), DMARDs treatment duration (in days), baseline HAQ score and their functional interactions with disease duration from symptom onset (in months).

The first scientific question is which of the aforementioned baseline covariates are good predictors of the disease progression outcome Y at the end of a 5-year follow-up. Second, as patients have various disease duration at registration (baseline), it is of interest to know whether the effects of the selected predictors on the outcome Y have interactions with the change of disease duration from disease symptom onset to the baseline time when these covariates were measured. Our assumption is that such interactions may not be fully modelled by adding linear interaction terms, and this motivates us to consider a semi-varying coefficient logistic regression to allow non-linear interactions.

3. Methodology

3.1. A semi-varying coefficient multinomial logistic regression model

Suppose we have a sample $(y_i, U_i, x_{i1}, \cdots, x_{id_n}), i = 1, \cdots, n$, from $(y, U, x_1, \cdots, x_{d_n})$. y is a categorical disease outcome variable of S levels of risk; U is a given continuous variable, such as baseline disease duration in the motivating example in Section 2; and $x_j, j = 1, \cdots, d_n$ are either continuous or discrete covariates whose effects may be constant or vary with the level of U. Throughout this paper, without loss of generality, we assume $y \in \{1, \cdots, S\}$ and take level S as reference.

Assume the conditional probability that the ith subject belongs to the risk category s is $p_{si} = P(y_i = s \mid U_i, x_{i1}, \cdots, x_{id_n})$, where $i = 1, \cdots, n$ and $s = 1, \ldots, S$. To incorporate non-linear interactions between x_j and U into the modelling, we specify all p_{si}s through a semi-varying coefficient multinomial logistic regression, that is,

$$\begin{aligned} p_{si} &= \frac{\exp\left(\sum_{j=1}^{d_n} x_{ij}a_{sj}(U_i)\right)}{1 + \sum_{k=1}^{S-1} \exp\left(\sum_{j=1}^{d_n} x_{ij}a_{kj}(U_i)\right)}, \quad s = 1, \ldots, S-1, \\ p_{Si} &= \frac{1}{1 + \sum_{k=1}^{S-1} \exp\left(\sum_{j=1}^{d_n} x_{ij}a_{kj}(U_i)\right)}, \end{aligned} \tag{3.1}$$

where $a_{kj}(\cdot)$ is unknown coefficients that are either constant or functional and $\sum_{s=1}^{S} p_{si} = 1$. A constant coefficient $a_{kj}(\cdot)$ means that there is no interaction between x_{ij} and U_i. It follows that the logit of category s versus the reference category S is $\ln\left(\frac{p_{si}}{p_{Si}}\right) = \sum_{j=1}^{d_n} x_{ij}a_{sj}(U_i)$.

3.2. Model selection, estimation and prediction

Throughout this section, for any function $f(\cdot)$, we use $\dot{f}(\cdot)$ to denote its first derivative and $\ddot{f}(\cdot)$ to denote its second derivative.

3.2.1. Model selection. We now describe how to select the predictor variables in (3.1) and identify which coefficients are constant and which are functional. This is basically a model selection problem. Based on the penalized likelihood idea, the model selection problem is transformed to an estimation problem of the unknown coefficients, $a_{kj}(\cdot)$s, in (3.1). In the following, we are going to apply the penalized local maximum likelihood estimation to estimate $a_{kj}(\cdot)$s in (3.1).

It is easy to see the conditional log-likelihood function of $a_{kj}(\cdot)$s, given all potential predictors, in (3.1) is

$$\sum_{i=1}^{n}\left\{\sum_{s=1}^{S-1} I(y_i=s)\sum_{j=1}^{d_n} x_{ij}a_{sj}(U_i) - \log\left(1+\sum_{l=1}^{S-1}\exp\left\{\sum_{j=1}^{d_n} x_{ij}a_{lj}(U_i)\right\}\right)\right\} \tag{3.2}$$

For each given k, $k=1,\dots,n$, within a small neighbourhood of U_k, a Taylor's expansion gives

$$a_{sj}(U_i) \approx a_{sj}(U_k) + \dot{a}_{sj}(U_k)(U_i - U_k),$$

where $i=1,\dots,n$, and $j=1,\dots,d_n$. This leads to the following local conditional log-likelihood function

$$\begin{aligned}\ell_k(\mathbf{a}_k,\ \mathbf{d}_k) = \sum_{i=1}^{n} K_h(U_i-U_k)\Bigg\{&\sum_{s=1}^{S-1} I(y_i=s)\sum_{j=1}^{d_n} x_{ij}\left\{\alpha_{sjk}+\beta_{sjk}(U_i-U_k)\right\}\\ &-\log\left(1+\sum_{l=1}^{S-1}\exp\left[\sum_{j=1}^{d_n} x_{ij}\left\{\alpha_{ljk}+\beta_{ljk}(U_i-U_k)\right\}\right]\right)\Bigg\}\end{aligned}$$

where α_{sjk} corresponds to $a_{sj}(U_k)$ and β_{sjk} corresponds to $\dot{a}_{sj}(U_k)$, $K(\cdot)$ is a kernel function, h is a bandwidth, $K_h(\cdot)=\frac{1}{h}K(\cdot/h)$,

$$\begin{aligned}\mathbf{a}_k &= \left(\alpha_{11k},\ \dots,\ \alpha_{1d_nk},\ \dots,\ \alpha_{(S-1)1k},\ \dots,\ \alpha_{(S-1)d_nk}\right)^{\mathrm{T}},\\ \mathbf{d}_k &= \left(\beta_{11k},\ \dots,\ \beta_{1d_nk},\ \dots,\ \beta_{(S-1)1k},\ \dots,\ \beta_{(S-1)d_nk}\right)^{\mathrm{T}}.\end{aligned}$$

Adding all $\ell_k(\mathbf{a}_k,\ \mathbf{d}_k)$, $k=1,\ \dots,\ n$, together, we have

$$\mathscr{L}_n(\mathscr{A},\ \mathscr{B}) = \sum_{k=1}^{n}\ell_k\left(\mathbf{a}_k,\ \mathbf{d}_k\right), \tag{3.3}$$

where

$$\mathscr{A} = \left(\mathbf{a}_1^{\mathrm{T}},\ \dots,\ \mathbf{a}_{\mathrm{n}}^{\mathrm{T}}\right)^{\mathrm{T}},\quad \mathscr{B} = \left(\mathbf{d}_1^{\mathrm{T}},\ \dots,\ \mathbf{d}_{\mathrm{n}}^{\mathrm{T}}\right)^{\mathrm{T}}.$$

Denote

$$\begin{aligned}&\|\mathbf{u}\| = \left(\mathbf{u}^{\mathrm{T}}\mathbf{u}\right)^{1/2},\quad \boldsymbol{\alpha}_{sj} = \left(\alpha_{sj1},\ \dots,\ \alpha_{sjn}\right)^{\mathrm{T}},\\ &\mathscr{D}_{sj} = \left\{\sum_{k=1}^{n}\left(\alpha_{sjk}-\bar{\alpha}_{sj}\right)^2\right\}^{1/2},\quad \text{and } \bar{\alpha}_{sj} = \frac{1}{n}\sum_{k=1}^{n}\alpha_{sjk}.\end{aligned}$$

This leads to the following penalized local conditional log-likelihood function for the model selection

$$\mathscr{Q}_n(\mathscr{A},\ \mathscr{B}) = \mathscr{L}_n(\mathscr{A},\ \mathscr{B}) - \sum_{s=1}^{S-1}\sum_{j=1}^{d_n} p_{\lambda_{1sj}}\left(\mathscr{D}_{sj}\right) - \sum_{s=1}^{S-1}\sum_{j=1}^{d_n} p_{\lambda_{2sj}}\left(\left\|\boldsymbol{\alpha}_{sj}\right\|\right), \tag{3.4}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function with tuning parameter λ, which is defined through its derivative

$$\dot{p}_\lambda(z) = \lambda\left[I(z \leqslant \lambda) + \frac{(a_0\lambda - z)_+}{(a_0-1)\lambda} I(z > \lambda)\right],$$

where $a_0 = 3.7$ as suggested in Fan and Li [22].

To directly maximize $\mathscr{Q}_n(\mathscr{A}, \mathscr{B})$ can be very challenging. We are going to find a quadratic function and use its maximizer to approximate the maximizer of $\mathscr{Q}_n(\mathscr{A}, \mathscr{B})$, thereby simplifying the maximization.

Let $(\widetilde{\mathscr{A}_n}, \tilde{\mathscr{B}}_n)$ be the maximizer of $\mathscr{L}_n(\mathscr{A}, \mathscr{B})$, $\tilde{\alpha}_{sjk}$ the component of $\widetilde{\mathscr{A}_n}$ which corresponds to α_{sjk}. $\tilde{\boldsymbol{\alpha}}_{sj}$ is $\boldsymbol{\alpha}_{sj}$ with α_{sjk} replaced by $\tilde{\alpha}_{sjk}$. $\tilde{\mathscr{D}}_{sj}$ is $\mathscr{D}_{sj}$ with α_{sjk} replaced by $\tilde{\alpha}_{sjk}$.

Noticing $\dot{\mathscr{L}}_n(\widetilde{\mathscr{A}_n}, \tilde{\mathscr{B}}_n) = 0$, by the Taylor's expansion, we have

$$\begin{aligned}\mathscr{L}_n(\mathscr{A}, \mathscr{B}) &\approx \mathscr{L}_n\left(\widetilde{\mathscr{A}_n}, \tilde{\mathscr{B}}_n\right) \\ &+ \frac{1}{2}\left(\left(\mathscr{A} - \widetilde{\mathscr{A}_n}\right)^{\mathrm{T}}, h\left(\mathscr{B} - \tilde{\mathscr{B}}_n\right)^{\mathrm{T}}\right) \ddot{\mathscr{L}}_n\left(\widetilde{\mathscr{A}_n}, \tilde{\mathscr{B}}_n\right) \begin{pmatrix} \mathscr{A} - \widetilde{\mathscr{A}_n} \\ h\left(\mathscr{B} - \tilde{\mathscr{B}}_n\right)\end{pmatrix},\end{aligned}$$

and for $s = 1, \cdots, S-1, j = 1, \cdots, d_n$,

$$p_{\lambda_{1sj}}\left(\mathscr{D}_{sj}\right) \approx p_{\lambda_{1sj}}\left(\tilde{\mathscr{D}}_{sj}\right) - \dot{p}_{\lambda_{1sj}}\left(\tilde{\mathscr{D}}_{sj}\right)\tilde{\mathscr{D}}_{sj} + \dot{p}_{\lambda_{1sj}}\left(\tilde{\mathscr{D}}_{sj}\right)\mathscr{D}_{sj},$$

$$p_{\lambda_{2sj}}\left(\left\|\boldsymbol{\alpha}_{sj}\right\|\right) \approx p_{\lambda_{2sj}}\left(\left\|\tilde{\boldsymbol{\alpha}}_{sj}\right\|\right) - \dot{p}_{\lambda_{2sj}}\left(\left\|\tilde{\boldsymbol{\alpha}}_{sj}\right\|\right)\left\|\tilde{\boldsymbol{\alpha}}_{sj}\right\| + \dot{p}_{\lambda_{2sj}}\left(\left\|\tilde{\boldsymbol{\alpha}}_{sj}\right\|\right)\left\|\boldsymbol{\alpha}_{sj}\right\|.$$

Let

$$\mathscr{L}_{n*}(\mathscr{A}, \mathscr{B}) = \frac{1}{2}\left(\left(\mathscr{A} - \widetilde{\mathscr{A}_n}\right)^{\mathrm{T}}, h\left(\mathscr{B} - \tilde{\mathscr{B}}_n\right)^{\mathrm{T}}\right) \ddot{\mathscr{L}}_n\left(\widetilde{\mathscr{A}_n}, \tilde{\mathscr{B}}_n\right) \begin{pmatrix} \mathscr{A} - \widetilde{\mathscr{A}_n} \\ h\left(\mathscr{B} - \tilde{\mathscr{B}}_n\right)\end{pmatrix},$$

and

$$\mathscr{P}_{1n,sj}\left(\mathscr{D}_{sj}\right) = \dot{p}_{\lambda_{1sj}}\left(\tilde{\mathscr{D}}_{sj}\right)\mathscr{D}_{sj}, \quad \mathscr{P}_{2n,sj}\left(\left\|\boldsymbol{\alpha}_j\right\|\right) = \dot{p}_{\lambda_{2sj}}\left(\left\|\tilde{\boldsymbol{\alpha}}_{sj}\right\|\right)\left\|\boldsymbol{\alpha}_{sj}\right\|$$

We define

$$\mathscr{Q}_{n*}(\mathscr{A}, \mathscr{B}) = \mathscr{L}_{n*}(\mathscr{A}, \mathscr{B}) - \sum_{s=1}^{S-1}\sum_{j=1}^{d_n}\mathscr{P}_{1n,sj}\left(\mathscr{D}_{sj}\right) - \sum_{s=1}^{S-1}\sum_{j=1}^{d_n}\mathscr{P}_{2n,sj}\left(\left\|\boldsymbol{\alpha}_{sj}\right\|\right),$$

and use the maximizer of $\mathscr{Q}_{n*}(\mathscr{A}, \mathscr{B})$ to approximate the maximizer of $\mathscr{Q}_n(\mathscr{A}, \mathscr{B})$ and estimate the corresponding unknown parameters.

By the local linear approximation, the maximization of $\mathscr{Q}_{n*}(\mathscr{A}, \mathscr{B})$ can be considered as an iterative reweighted LASSO problem. Hence given an initial estimator, $\mathscr{Q}_{n*}(\mathscr{A}, \mathscr{B})$ can be maximized through an iterative algorithm similar as in Li *et al.* [28]. The computational cost is moderate.

Let $\left(\hat{\boldsymbol{\alpha}}_{sj}, \hat{\boldsymbol{\beta}}_{sj}\right)$, $s = 1, \cdots, S-1, j = 1, \cdots, d_n$, be the maximizer of $\mathscr{Q}_{n*}(\mathscr{A}, \mathscr{B})$. For the penalty functions which enjoy sparsity property, such as SCAD or L_1 penalty, our feature selection and model specification procedure work as follows: if $\|\hat{\boldsymbol{\alpha}}_{sj}\| = 0$, then the corresponding variable x_j is not significant and should be removed from modelling the conditional probability $P\left(y = s|U, x_1, \cdots, x_{d_n}\right)$ of y falling in level s. Let $\hat{\mathscr{D}}_{sj}$ be $\mathscr{D}_{sj}$ with $\boldsymbol{\alpha}_{sj}$ replaced by $\hat{\boldsymbol{\alpha}}_{sj}$. If $\hat{\mathscr{D}}_{sj} = 0$, the coefficient of x_j is constant when modelling $P\left(y = s|U, x_1, \cdots, x_{d_n}\right)$.

3.2.2. *Estimation.*

After the model is selected, we apply the standard local maximum likelihood estimation to estimate the coefficients based on the selected model. The details are as follows.

Suppose the set of the subscripts of the variables with functional coefficients, in the selected model for $P(y = s|U, x_1, \cdots, x_{d_n})$, is Ω_s, with constant coefficients is Δ_s. For any given u, by simple calculation, we have the following local conditional log-likelihood function:

$$\begin{aligned}\sum_{i=1}^{n} K_h(U_i - u)\left\{\sum_{s=1}^{S-1} I(y_i = s)\left[\sum_{j\in\Omega_s} x_{ij}\left\{\alpha_{sj} + \beta_{sj}(U_i - u)\right\} + \sum_{l\in\Delta_s} x_{il}\alpha_{sl}\right]\right. \\ \left. - \log\left(1 + \sum_{k=1}^{S-1}\exp\left[\sum_{j\in\Omega_k} x_{ij}\left\{\alpha_{kj} + \beta_{kj}(U_i - u)\right\} + \sum_{l\in\Delta_k} x_{il}\alpha_{kl}\right]\right)\right\}.\end{aligned}$$

Let $(\hat\alpha_{sj}(u),\ \hat\beta_{sj}(u)),\ j \in \Omega_s \cup \Delta_s,\ s = 1, \cdots, S-1$, be the maximizer of this local conditional log-likelihood function at u.

For any $j \in \Omega_s$, the estimator $\hat a_{sj}(u)$ of the functional coefficient $a_{sj}(u)$ is taken to be $\hat\alpha_{sj}(u)$. For any $l \in \Delta_s$, the coefficient $a_{sl}(\cdot)$ is constant, which is denoted by C_{sl}, and can be estimated by

$$\hat C_{sl} = \frac{1}{n}\sum_{i=1}^{n} \hat\alpha_{sl}(U_i).$$

Remark 1

The Taylor's expansion-based local maximum likelihood estimation used in this paper is an important nonparametric estimation method. Theoretically speaking, the estimation error due to the linear approximation is of order $O_p(h^2)$, see Li *et al.* [28]. Practically, this would be very small as long as the sample size is reasonable. The limitation of the local maximum likelihood estimation is that all functions, which are approximated by linear functions, are required to have continuous second derivative.

3.2.3. Prediction. Once the model is specified and the coefficients in the selected model are estimated, the risk prediction becomes straightforward: for a new subject, if the observation of the predictor is $\left(U_l,\ x_{l1},\ \dots,\ x_{ld_n}\right)$, the conditional probability of this subject falling in risk level s, $s \in \{1,\ \dots,\ S-1\}$, given $\left(U_l,\ x_{l1},\ \dots,\ x_{ld_n}\right)$ can be estimated by

$$\hat p_{sl} = \frac{\exp\left(\sum_{j\in\Omega_s} x_{lj}\hat a_{sj}(U_l) + \sum_{j\in\Delta_s} x_{lj}\hat C_{sj}\right)}{1 + \sum_{k=1}^{S-1}\exp\left(\sum_{j\in\Omega_k} x_{lj}\hat a_{kj}(U_l) + \sum_{j\in\Delta_k} x_{lj}\hat C_{kj}\right)}. \tag{3.5}$$

Let

$$\hat p_{Sl} = 1 - \sum_{s=1}^{S-1} p_{sl}$$

and $\hat s$ maximize $\hat p_{sl}$ with respect to s on $\{1,\ \dots,\ S\}$. If $\hat p_{\hat s l}$ is greater than a given threshold, this new subject is predictable under this threshold. Then, we classify it into risk level $\hat s$.

3.3. Selection of tuning parameters

The tuning parameters λ_{1sj} and $\lambda_{2sj}, j = 1, \cdots, d_n, s = 1, \cdots, S-1$, involved in the proposed estimation and model selection procedure play a very important role. In this section, we will address how to choose these tuning parameters.

If we use a proper penalty function, such as SCAD, it would be reasonable to set all λ_{1sj}'s to have the same value and all λ_{2sj}'s to have the same value. This is because the need of different tuning parameters for different coefficients would be met by the use of a proper penalty function. From now on, we use λ_1 and λ_2 to denote the common value of λ_{1sj}'s and λ_{2sj}'s, respectively, and select λ_1 and λ_2 by the generalized information criterion (GIC) proposed by Fan and Tang [40].

As the model concerned involves both unknown constant parameters and unknown functional parameters, to use GIC, we first need to figure out how many unknown constant parameters an unknown functional parameter amounts to. Cheng *et al.* [10] suggested that an unknown functional parameter would amount to 1.028571 h^{-1} unknown constant parameters when Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$ was used. Taking their suggestion, we construct the GIC for model (3.3) as

$$\begin{aligned}\mathrm{GIC}(\lambda_1,\ \lambda_2) = &-2\sum_{k=1}^{n} \ell_k\left(\hat{\mathbf{a}}_k,\ \hat{\mathbf{d}}_k\right),\\ &+ 2\ln\{\ln(n)\}\ln\left\{1.028571(S-1)d_n h^{-1}\right\}\left(k_1 + 1.028571 k_2 h^{-1}\right),\end{aligned}$$

where $\hat{\mathbf{a}}_k$ and $\hat{\mathbf{d}}_k$ are the estimators of $\mathbf{a}_k$ and $\mathbf{d}_k$ obtained based on the tuning parameters λ_1 and λ_2. k_1 is the number of significant covariates with constant coefficients obtained based on the tuning parameters

λ_1 and λ_2, and k_2 is the number of significant covariates with functional coefficients obtained based on the tuning parameters λ_1 and λ_2. The minimizer of GIC$(\lambda_1,\ \lambda_2)$ is the selected λ_1 and λ_2.

4. Application to the Health Assessment Questionnaire progression data

4.1. Data analysis

We fitted model (3.1) to the HAQ data and considered the disease duration variable as the covariate U. One of the advantages of the proposed semi-varying coefficient multinomial logistic regression model is that it allows us to incorporate potentially varying effects of baseline covariates with the change of U. It is more flexible and more general than the one including linear interaction terms between baseline covariates and U. Without loss of generality, we rescaled the covariate U to [0, 1]. The response category $Y = 1$ was chosen as the reference, and the other two categories were compared against the reference category. We considered 18 candidate covariates including all numerical or dummy variables listed in Section 2.2. We used the proposed risk prediction methods discussed in Section 3.2 to do the model selection, estimation and prediction. The kernel function was chosen as the Epanechnikov kernel, and the bandwidth was chosen as $h = 0.6[(S-1)d_n/n]^{0.2}$. The tuning parameters λ_1 and λ_2 were selected by GIC [28, 40] as discussed in Section 3.3.

The selected predictors together with their estimated coefficients (constant or functional) are presented in Tables I and II. For those functional coefficients, we report their estimates and associated standard errors at given U values with the disease duration being 1, 3, 6, 12 or 24 months. The standard errors of the coefficient estimates were calculated by a bootstrap method. Among the list of candidate covariates, 12 were selected to be significantly associated with the multinomial logit of the response group $Y = 2$ (moderate risk) relative to the reference group $Y = 1$ (low risk). Three of them (RA, female and current smoker) have constant coefficients and are associated with increased probability of being in the higher risk groups. The others (baseline HAQ score, number of swollen joints, number of tender joints, DMARDs treatment duration, age at onset, copies of genetic biomarker, previous smoker, upper middle deprived group and most deprived group) have functional coefficients, which mean that their effects are

Table I. HAQ progression data: selected covariates for the logit of $Y = 2$ (moderate risk) vs $Y = 1$ (low risk) and their coefficient estimates (standard errors) that are either constant or varying at selected disease durations.

Variable	Type		1 month	3 months	6 months	12 months	24 months
RA	Constant	0.091	—	—	—	—	—
(yes or no)		(0.051)					
Female	Constant	0.144	—	—	—	—	—
(yes or no)		(0.010)					
Current smoker	Constant	0.093	—	—	—	—	—
(yes or no)		(0.002)					
Baseline HAQ score	Functional	—	0.147	0.084	0.044	0.089	0.576
		—	(0.041)	(0.019)	(0.017)	(0.023)	(0.153)
Number of swollen joints	Functional	—	−0.011	−0.005	0.001	0.001	−0.028
		—	(0.004)	(0.002)	(0.0003)	(0.0004)	(0.007)
Number of tender joints	Functional	—	−0.007	−0.004	−0.0004	0.007	0.013
		—	(0.003)	(0.001)	(0.0001)	(0.002)	(0.005)
Treatment duration	Functional	—	−0.0002	−0.0001	0.0001	0.0002	0.0002
(days)		—	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Age at onset	Functional	—	−0.002	−0.001	−0.0004	0.000	−0.0002
		—	(0.0005)	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Copies of genetic	Functional	—	0.078	0.054	0.038	0.006	−0.104
biomarker		—	(0.028)	(0.021)	(0.015)	(0.002)	(0.033)
Previous smoker	Functional	—	0.107	0.080	0.030	−0.032	−0.173
(yes or no)		—	(0.027)	(0.019)	(0.009)	(0.007)	(0.057)
Upper middle deprived	Functional	—	0.291	0.241	0.194	0.083	−0.184
group (yes or no)		—	(0.086)	(0.089)	(0.046)	(0.032)	(0.053)
Most deprived group	Functional	—	0.309	0.290	0.264	0.174	−0.176
(yes or no)		—	(0.101)	(0.093)	(0.088)	(0.067)	(0.068)

RA, fulfilment of the ACR 1987 classification criteria for rheumatoid arthritis; HAQ, Health Assessment Questionnaire.

Table II. HAQ progression data: selected covariates for the logit of $Y = 3$ (high risk) vs $Y = 1$ (low risk) and their coefficient estimates (standard errors) that are either constant or varying at selected disease durations.

Variable	Type		1 month	3 months	6 months	12 months	24 months
Baseline HAQ score	Constant	0.052	—	—	—	—	—
		(0.0004)					
Rheumatic factor	Constant	0.077	—	—	—	—	—
(positive or negative)		(0.0008)					
Intercept	Functional	—	−0.159	−0.149	−0.138	−0.098	0.025
		—	(0.043)	(0.039)	0.052)	(0.034)	(0.007)
Number of swollen joints	Functional	—	0.005	0.004	0.003	0.004	0.010
		—	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)
Number of tender joints	Functional	—	−0.006	−0.004	−0.001	0.002	0.015
		—	(0.002)	(0.001)	(0.0003)	(0.001)	(0.007)
Treatment duration	Functional	—	0.001	0.0004	0.0001	−0.0002	−0.0001
(days)		—	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0001)
Age at onset	Functional	—	0.003	0.002	0.002	0.001	−0.001
		—	(0.001)	(0.001)	(0.0004)	(0.0002)	(0.0003)
Copies of genetic	Functional	—	0.057	0.053	0.038	0.003	−0.054
biomarker		—	(0.018)	(0.016)	(0.015)	(0.001)	(0.021)
Upper middle deprived	Functional	—	−0.095	−0.045	0.020	0.117	0.165
group (yes or no)		—	(0.019)	(0.016)	(0.008)	(0.044)	(0.038)

HAQ, Health Assessment Questionnaire.

varying with the change of baseline disease duration. For the multinomial logit of the response group $Y = 3$ (high risk) relative to the low risk group, eight covariates together with a functional intercept were selected in the model. Two of them (baseline HAQ and rheumatic factor) have constant coefficients, and six (number of swollen joints, number of tender joints, DMARDs treatment duration, age at onset, copies of genetic biomarker and upper middle deprived group) have functional coefficients. All of the selected covariates in Tables I and II are indeed well acknowledged predictors in HAQ progression (for example, Combe *et al.* [41]). Fewer predictors were identified in Table II because of the smaller sample size of 23 in Group $Y = 3$ comparing with 74 in Group $Y = 2$. It is interesting but not surprising to see that the effects of smoking status and genetic biomarker with short baseline disease duration (1 or 3 months) contribute more toward higher risk, while the effects of clinical variables (i.e. baseline HAQ sore and number of tender joints) with long baseline disease duration (24 months) contribute more toward higher risk.

The scientific aim of this study is to classify the patients at baseline into different risk groups to predict their outcomes at the end of follow-up. Hence, we assess the performance of our methods by comparing the correct prediction rate with other existing methods. The calculation of the correct prediction rate is based on a leave-one-out cross-validation approach. For each subject, we used the rest of the data (289 subjects) to select covariates and obtain their coefficient estimates. We then calculated the estimated conditional probability of belonging to low, moderate or high-risk levels for this subject. If one of the estimated conditional group-membership probabilities is higher than a threshold, say 80% or 70%, we classify this subject into the corresponding group and compare the prediction result with the true value of Y. By repeating this procedure to all subjects, we calculated correct prediction rates for those subjects who have a maximum of the estimated group-membership probabilities greater than the threshold. The results are shown in Table III, where the ‘Total prediction No.’ means the numbers of subjects with a maximum group-membership probability greater than the threshold (80% or 70%) and the ‘Correct prediction No.’ means the numbers of subjects correctly grouped. The correct prediction rates, which are ratios between the correct prediction numbers and the total prediction numbers, are compared between our method and alternative competitors. The alternative competitors include ‘SCAD constant coefficient model’, ‘Full varying coefficient model’ and ‘Full constant coefficient model’. The ‘SCAD constant coefficient model’ is selected by a SCAD penalized multinomial logistic regression with constant coefficients. The ‘Full varying coefficient model’ means a varying coefficient multinomial logistic regression including all covariates with functional coefficients. The ‘Full constant coefficient model’ means a multinomial logistic regression including all covariates with constant coefficients. We see that our selected model always gives better correct prediction rates comparing with all the competitors by reducing false prediction numbers significantly. Its correct prediction rate could reach as high as 85.2% when the threshold probability is 0.8. The SCAD constant coefficient model performs better than the two full models.

Table III. HAQ progression data: comparison of the total prediction numbers, the correct prediction numbers and the correct prediction rates among models.

Model	Total prediction no.	Correct prediction no.	Correct prediction rate (%)
		Estimated conditional probability ⩾ 80%	
Our method	61	52	85.2
SCAD constant coefficient model	72	51	70.8
Full varying coefficient model	77	51	66.2
Full constant coefficient model	81	50	61.7
		Estimated conditional probability ⩾ 70%	
Our method	120	93	77.5
SCAD constant coefficient model	130	89	68.5
Full varying coefficient model	138	88	63.7
Full constant coefficient model	143	90	62.9

HAQ, Health Assessment Questionnaire; SCAD, smoothly clipped absolute deviation.

The aforementioned results justified the motivation of the proposed method as both varying coefficient assumption and model selection are crucial to the improvement of prediction performance. The total computation time for this section is about 2 h.

4.2. *Reliability of prediction*

To evaluate the reliability of the proposed risk-prediction model, we consider a recalibration framework in this section. Recalibration methods that involve the estimation of calibration intercept and calibration slope are well recognized approaches to assess the reliability of prediction models [42, 43]. The logistic recalibration framework for the risk-prediction models with binary outcomes has been proposed in existing literature [44, 45]. Recently, Van Hoorde *et al.* [46] extended this recalibration tool for nominal polytomous outcomes. In this paper, we extend the recalibration framework proposed by Van Hoorde *et al.* [46] to propose new calibration tools, including calibration intercept and calibration slope, to assess the reliability of the proposed prediction model based on semi-varying coefficient multinomial logistic regression. The calibration intercept, calibration slope and their confidence intervals were calculated as follows.

We divided the HAQ data sample into two datasets: a prediction dataset including 200 subjects and a validation dataset including 90 subjects. First, we used the prediction data to do the model selection. Based on the selected model and the estimated coefficients, we used the validation data to estimate the intercept and the slope of a recalibration multinomial regression model as follows.

Denote Ω_s and Δ_s the set of the subscripts of the variables with functional and constant coefficients in the selected model for category s, respectively, $s = 2, 3$. We fitted the following recalibration model to the validation data for $l = 1, \ldots, 90$:

$$\begin{cases} \log\left[\frac{P(Y_l=2)}{P(Y_l=1)}\right] = \gamma_{20} + \gamma_{21}\left(\sum_{j\in\Omega_2} x_{lj}\hat{a}_{2j}(U_l) + \sum_{j\in\Delta_2} x_{lj}\hat{C}_{2j}\right), & \text{for } Y_l = 2, \\ \log\left[\frac{P(Y_l=3)}{P(Y_l=1)}\right] = \gamma_{30} + \gamma_{31}\left(\sum_{j\in\Omega_3} x_{lj}\hat{a}_{3j}(U_l) + \sum_{j\in\Delta_3} x_{lj}\hat{C}_{3j}\right), & \text{for } Y_l = 3, \end{cases} \tag{4.1}$$

where $\hat{a}_{2j}(.)$, $\hat{a}_{3j}(.)$, $\hat{C}_{2j}$ and $\hat{C}_{3j}$ were obtained by applying the method proposed in Section 3.2.2 to the prediction data; γ_{s0} and γ_{s1} are the intercept and the slope for the sth category, $s = 2, 3$. We estimated the calibration slopes $\hat{\gamma}_{s1}$ and calculated their 95% Wald-type confidence intervals. Similarly, we calculated the calibration intercepts and their 95% Wald-type confidence intervals except that all $\hat{\gamma}_{s0}$s were obtained by fixing the corresponding slopes γ_{s1} equal to 1 [46].

The calibration slope and the calibration intercept are important aspects of calibration, assessing the amount of model overfitting and difference between observed outcome versus predicted outcome, respectively [42, 43]. When the risks are perfectly calibrated, the calibration slope is 1, and the calibration intercept is 0 [46]. In addition, their confidence intervals should be reasonably small. Comparisons of the estimated calibration slopes and calibration intercepts between our method, the ‘SCAD constant coef-

Table IV. HAQ progression data: estimated calibration intercepts and calibration slopes together with their 95% Wald-type confidence intervals (in the brackets); the category $Y = 1$ is the reference.

	Calibration intercept	Calibration slope
Our method		
$Y = 2$	−0.16 [−0.22; −0.06]	1.06 [0.93; 1.23]
$Y = 3$	0.23 [0.07; 0.38]	1.08 [0.88; 1.32]
SCAD constant coefficient model		
$Y = 2$	−0.55 [−0.86; −0.22]	0.83 [0.70; 1.09]
$Y = 3$	0.44 [0.15; 0.81]	0.86 [0.72; 1.13]
Full varying coefficient model		
$Y = 2$	−0.65 [−0.84; −0.34]	0.76 [0.47; 1.07]
$Y = 3$	1.04 [0.66; 1.32]	0.78 [0.53; 1.04]

HAQ, Health Assessment Questionnaire; SCAD, smoothly clipped absolute deviation.

ficient model' and the 'Full varying coefficient model' are presented in Table IV. It shows clearly that the model selected by our method is more reliable because its calibration intercept is more close to 0 (observed outcomes agree more with predicted risks) and its calibration slope is more close to 1. The calibration slopes for the 'SCAD constant coefficient model' (i.e. 0.83 and 0.86) and for the 'Full varying coefficient model' (i.e. 0.76 and 0.78) suggest more overfitting in both models. The total computation time for this section is about 1 h.

4.3. Simulation study

To examine the performance of the proposed method in Section 3, we conduct a simulation study to answer two questions: (i) whether the proposed model selection procedure works well to select correct predictors and the types of their coefficients (constant or functional); (ii) whether the developed prediction model works well to classify subjects into different risk groups using baseline information.

We generated a simulated dataset that mimics the real HAQ progression data. The covariates $\mathbf{x}_j$'s were generated independently as follows. If the jth covariate in the HAQ progression data is continuous, $\mathbf{x}_j$ was generated from a normal distribution with its mean and variance being equal to the sample mean and the sample variance of the jth covariate. If the jth covariate is discrete, then $\mathbf{x}_j$ was generated from a discrete distribution taking the same values with probabilities being equal to the sample probabilities estimated from the HAQ progression data. The covariate U was generated independently from *Uniform* $[0,\ 1]$. The response variable $Y \in \{1,\ 2,\ 3\}$ was generated from the following multinomial logistic regression model:

$$\begin{cases} \log\left[\frac{P(Y=2)}{P(Y=1)}\right] = \sum\limits_{j\in\Omega_2} x_j a_{2j}(U) + \sum\limits_{j\in\Delta_2} x_j C_{2j}, & \text{for } Y=2, \\ \log\left[\frac{P(Y=3)}{P(Y=1)}\right] = \sum\limits_{k\in\Omega_3} x_k a_{3j}(U) + \sum\limits_{k\in\Delta_3} x_k C_{3j}, & \text{for } Y=3, \end{cases} \tag{4.2}$$

where Ω_2, Ω_3, Δ_2 and Δ_3 refer to the model selected in Section 4.1; $a_{2j}(\cdot)$, $a_{3j}(\cdot)$, C_{2j} and C_{3j} are the same as the estimated coefficients in Section 4.1. In Section 4.3, model (4.2) is considered as the true model and $a_{2j}(\cdot)$, $a_{3j}(\cdot)$, C_{2j} and C_{3j} are considered as the true coefficients.

We considered different scenarios with sample sizes n being equal to 200, 300 or 400. For each scenario, we generated a training dataset of sample size n from (4.2). The simulation was conducted in the following steps:

(i) Select the model and estimate $a_{2j}(\cdot)$, $a_{3j}(\cdot)$, C_{2j} and C_{3j} using the training data;
(ii) Generate a separate testing dataset such that there are 100 subjects in each response category; apply the estimated model obtained in Step (i) to predict the response Y for each subject in this testing dataset;
(iii) Generate a second testing dataset such that there is one response category into which subjects are classified with conditional probabilities of 90–100%, 80–90% and 70–80%, respectively; keep simulating until each response category has a sample size of 100; apply the estimated model obtained in Step (*i*) to predict the response Y for each subject in the second testing dataset.

By repeating the aforementioned procedure for 200 realizations, we assess the model selection performance by reporting the average numbers of correct-selection, wrong-selection and missing-selection for the logit of $Y = 2$ vs $Y = 1$ and the logit of $Y = 3$ vs $Y = 1$, respectively. The results are presented in Table V, where 'Correct-selection' means that the selected covariates are true predictors, and the procedure correctly selects their coefficient types (constant or functional); 'Wrong-selection' means that some selected covariates are false predictors or the procedure wrongly selects their coefficient types; 'Missing-selection' means that some true predictors are not selected in the procedure. It is seen that our proposed method performs well in the model selection with small average numbers of wrong-selection and missing-selection. For example, when selecting 13 (or 10) true predictors for the logit of $Y = 2$ (or $Y = 3$) vs $Y = 1$ among 21 candidate covariates, there is about one covariate wrongly selected on average for $n = 300$, mainly because of wrong-selection for its coefficient type. It is less likely to miss a true predictor with an average number of missing-selection being 0.23 (or 0.20) for $n = 300$. When the sample size is reduced to 200, the average numbers of wrong-selection and missing-selection slightly increase to 2 and 1, respectively.

In addition, we assess the out-of-sample performance of the estimated prediction model by applying it to separate testing datasets. The average proportions of correct prediction in each category in Step (ii) and the average proportions of subjects being classified into a correct category in Step (iii) are reported in

Table V. Simulation study: comparison of the performances of our method, the SCAD constant coefficient model and the full constant coefficient model.

	(i) Average number of correct, wrong or missing-selection			
Sample size	Category	Correct-selection	Wrong-selection	Missing-selection
$n = 200$	$Y = 2$	10.35	2.12	1.63
	$Y = 3$	7.95	1.82	1.05
$n = 300$	$Y = 2$	11.77	1.12	0.23
	$Y = 3$	8.80	1.06	0.20
$n = 400$	$Y = 2$	11.90	0.95	0.10
	$Y = 3$	8.86	0.91	0.14
	(ii) Average proportion of correct prediction within each category			
Sample size	Model	$Y = 1$	$Y = 2$	$Y = 3$
$n = 200$	Our method	0.68	0.62	0.56
	SCAD constant coefficient model	0.55	0.53	0.50
	Full constant coefficient model	0.31	0.30	0.27
	Oracle model	0.79	0.74	0.67
$n = 300$	Our method	0.74	0.69	0.60
	SCAD constant coefficient model	0.55	0.53	0.50
	Full constant coefficient model	0.38	0.36	0.32
	Oracle model	0.79	0.74	0.67
$n = 400$	Our method	0.78	0.75	0.71
	SCAD constant coefficient model	0.63	0.60	0.59
	Full constant coefficient model	0.44	0.41	0.38
	Oracle model	0.79	0.74	0.67
	(iii) Average proportion of subjects being classified into a correct category			
Sample size	Model	90–100%	80–90%	70–80%
$n = 200$	Our method	0.88	0.76	0.65
	SCAD constant coefficient model	0.74	0.62	0.51
	Full constant coefficient model	0.69	0.54	0.41
	Oracle model	0.94	0.83	0.73
$n = 300$	Our method	0.92	0.80	0.69
	SCAD constant coefficient model	0.77	0.68	0.56
	Full constant coefficient model	0.72	0.58	0.45
	Oracle model	0.94	0.83	0.73
$n = 400$	Our method	0.93	0.82	0.70
	SCAD constant coefficient model	0.78	0.70	0.59
	Full constant coefficient model	0.74	0.60	0.49
	Oracle model	0.94	0.83	0.73

SCAD, smoothly clipped absolute deviation.

the Panels (ii) and (iii) of Table V, respectively. In order to make a comparison, we report the simulation results from the 'Oracle model', which means that we know the true model and the true coefficients, and a number of competitor methods available to be implemented, including the 'SCAD constant coefficient model' and the 'Full constant coefficient model', as explained in Section 4.1. We see that the prediction performance of the estimated model is obviously better than those the two competitor methods and is close to that of the 'Oracle model' when $n = 400$. The computation of the simulation was performed using a small number of computer clusters, and the execution time is about 8 h. Additional simulation results are given in online supplementary materials.

5. Discussion

In the management of chronic diseases, it is of interest to identify a subgroup of subjects at baseline who are at high risk in future progression to a severe disease outcome, and hence, specific therapeutic strategies could be matched. Many prognostic markers (predictors) are often taken into account in risk-prediction modelling, and the interactions between predictor variables can be complicated. In this paper, we presented a semi-varying coefficient regression model for improving the prediction modelling and conducted the model selection by a penalized likelihood approach. Based on the ideas of penalization on deviation, kernel smoothing and quadratic function approximation, our method selects significant predictors, determines whether each selected predictor has a constant or functional coefficient and estimates their coefficients simultaneously. Another attractive feature of the proposed method is that it allows the number of potential covariates to increase with the sample size. With rapid development of laboratory medicine, more potential prognostic markers, including clinical and demographic features, environmental factors, serological factors, genetic factors, epigenetic factors and their interactions, are considered as candidates predicting future disease outcome or response to treatment in stratified medicine, and the number of potential predictors could be very large. This paper focuses on nonparametric risk prediction modelling, and future work would be focussing on treatment-specific consideration in stratified medicine and methods to predict response to treatment.

We consider the multinomial logistic regression as it is widely used in medical research for categorical outcomes. It would be a topic of future work to extend the proposed method to an ordinal logistic regression context for the ordinal outcome HAQ. For example, one could extend the current method by penalizing the likelihood of a proportional odds logistic regression [47] for ordinal outcomes in a similar way. However, a different algorithm may be required to handle the optimization of its penalized likelihood with constrained parameterization.

Our model allows non-linear interactions that include the usual type of interactions, such as product of x and U, as special cases. For example, if the function $a(U)$ in model (3.1) is linear, such as $a(U)x = aUx$, the model is specialized to one with an usual type of linear interaction. From a modelling point of view, one can also entertain other potential interactions between other covariates by including extra non-linear interaction terms (or other transformation) of the covariates concerned in the model.

Acknowledgements

We are very grateful to the editor, the associate editor and two anonymous referees for their helpful comments. An earlier version of this paper was submitted when Bo Fu worked in the University of Manchester, and we wish to thank the PI in the Arthritis Research UK Epidemiology Unit for allowing us to use the NOAR data. Bo Fu was partially supported by the UK Medical Research Council research grant (MR/M025152/1) and a key international collaboration project (71420107023) funded by National Natural Science Foundation of China.

References

1. Vastesaeger N, Xu S, Aletaha D, St Clair E, Smolen J. A pilot risk model for the prediction of rapid radiographic progression in rheumatoid arthritis. *Rheumatology* 2009; **48**(9):1114–1121.
2. Visser K, Goekoop-Ruiterman Y, de Vries-Bouwstra J, Ronday H, Seys P, Kerstens P, et al. A matrix risk model for the prediction of rapid radiographic progression in patients with rheumatoid arthritis receiving different dynamic treatment strategies: post hoc analyses from the best study. *Annals of the Rheumatic Diseases* 2010; **69**(7):1333–1337.
3. Norton S, Fu B, Scott D, Deighton C, Symmons D, Wailoo A, Tosh J, Lunt M, Davies R, Young A, Verstappen S. Health Assessment Questionnaire disability progression in early rheumatoid arthritis: systematic review and analysis of two inception cohorts. *Seminars in Arthritis and Rheumatism* 2014; **44**(2):131–144.
4. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. John Wiley & Sons: New York, 2004.

5. Li J, Jiang B, Fine J. Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. *Biostatistics* 2013; **14**(2):382–394.
6. Cleveland WS, Grosse E, Shyu WM. Local regression models. In *Statistical Models in S*, Chambers JM, Hastie T (eds). Wadsworth and Brooks/Cole: Pacific Grove, 1991; 309–376.
7. Fan J, Zhang W. Statistical estimation in varying coefficient models. *The Annals of Statistics* 1999; **27**:1491–1518.
8. Fan J, Zhang W. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics* 2000; **27**:715–731.
9. Sun Y, Zhang W, Tong H. Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics* 2007:2795–2814.
10. Cheng M, Zhang W, Chen L. Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association* 2009; **104**:1179–1191.
11. Wang H, Xia Y. Shrinkage estimation of the varying-coefficient model. *Journal of the American Statistical Association* 2009; **104**:747–757.
12. Zhang W, Fan J, Sun Y. A semiparametric model for cluster data. *The Annals of Statistics* 2009; **37**:2377–2408.
13. Wu Y, Fan J, Muller H. Varying-coefficient functional linear regression. *Bernoulli* 2010; **16**:730–758.
14. Jiang J. Multivariate functional-coefficient regression models for multivariate nonlinear times series. *Biometrika* 2014; **101**(3):689–702.
15. Fan J, Zhang W. Statistical methods with varying coefficient models. *Statistics and its Interface* 2008; **1**(1):179–195.
16. Solari A, Saskia C, Jelle JG. Testing goodness of fit in regression: a general approach for specified alternatives. *Statistics in Medicine* 2012; **31**(28):3656–3666.
17. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. *International Journal of Epidemiology* 2007; **36**(1):195–202.
18. Breiman L. Heuristics of instability and stabilization in model selection. *The Annals of Statistics* 1996; **24**:2350–2383.
19. Donoho DL. High-dimensional data analysis: the curses and blessings of dimensionality. *AMS Math Challenges Lecture* 2000; 1–32.
20. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**:55–67.
21. Tibshirani RJ. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B* 1996; **58**:267–288.
22. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96**:1348–1360.
23. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B* 2006; **68**:49–67.
24. Zou H. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* 2006; **101**: 1418–1429.
25. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 2010; **38**:894–942.
26. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association* 1988; **83**:1023–1036.
27. Lian H. Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica* 2012; **22**: 1563–1588.
28. Li D, Ke Y, Zhang W. Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models. *The Annals of Statistics* 2015; **43**(6):2676–2705.
29. Kong E, Xia Y. Variable selection for the single-index model. *Biometrika* 2006; **94**:217–229.
30. Tao H, Xia Y. Adaptive semi-varying coefficient model selection. *Statistica Sinica* 2011; **22**:575–599.
31. Kuk AYC, Li J, Rush JA. Variable and threshold selection to control predictive accuracy in logistic regression. *Applied Statistics* 2014; **63**(4):657–672.
32. Stefanski L, Wu Y, White K. Variable selection in nonparametric classification via measurement error model selection likelihoods. *Journal of the American Statistical Association* 2014; **109**(506):574–589.
33. Farragher T, Lunt M, Fu B, Bunn D, Symmons D. Early treatment with, and time receiving, first disease-modifying antirheumatic drug predicts long-term function in patients with inflammatory polyarthritis. *Annals of the Rheumatic Diseases* 2010; **69**(4):689–695.
34. Visser H. Early diagnosis of rheumatoid arthritis. *Best Practice & Research Clinical Rheumatology* 2005; **19**:55–72.
35. Dixon N, Symmons D. Does early rheumatoid arthritis exist? *Best Practice & Research Clinical Rheumatology* 2005; **19**:37–53.
36. Combe B, Landewe R, Lukas C, Bolosiu HD, Breedveld F, Dougados M, Emery P, Ferraccioli G, Hazes JM, Klareskog L, Machold K, Martin-Mola E, Nielsen H, Silman A, Smolen J, Yazici H. EULAR evidence recommendations for the management of early arthritis. Report of a task force of the European standing committee for international clinical studies including therapeutics. *Annals of the Rheumatic Diseases* 2007; **66**:34–45.
37. Venkateshan SP, Sidhu S, Malhotra S, Pandhi P. Efficacy of biologicals in the treatment of rheumatoid arthritis: a meta-analysis. *Pharmacology* 2009; **83**:1–9.
38. Fu B, Lunt M, Galloway J, Dixon W, Hyrich K, Symmons D. A threshold hazard model for estimating serious infection risk following anti-tumor necrosis factor therapy in rheumatoid arthritis patients. *Journal of Biopharmaceutical Statistics* 2013; **23**(2):461–476.
39. Wolfe F, Kleinheksel SM, Cathey MA, Hawley DJ, Spitz PW, Fries JF. The clinical value of the stanford health assessment questionnaire functional disability index in patients with rheumatoid arthritis. *The Journal of Rheumatology* 1988; **15**: 1480–1488.
40. Fan Y, Tang CY. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society Series B* 2012; **75**:531–552.

41. Combe B, Cantagrel A, Goupille P, Bozonnat MC, Sibilia J, Eliaou JF, Meyer O, Sany J, Dubois A, Daurès JP, Dougados M. Predictive factors of 5-year health assessment questionnaire disability in early rheumatoid arthritis. *The Journal of Rheumatology* 2003; **30**:2344–2349.
42. Cox D. Two further applications of a model for binary regression. *Biometrika* 1958; **45**:562–565.
43. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. *Medical Decision Making* 1993; **13**(1):49–57.
44. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* 2010; **21**(1):128–138.
45. Viallon V, Ragusa S, Clavel-Chapelon F, Bénichou J. How to evaluate the calibration of a disease risk prediction tool. *Statistics in Medicine* 2009; **28**(6):901–916.
46. Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B. Assessing calibration of multinomial risk prediction models. *Statistics in Medicine* 2014; **33**(15):2585–2596.
47. Brant R. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* 1990; **46**(4): 1171–1178.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.