International Journal of Epidemiology, 2018, 1383–1388
doi: 10.1093/ije/dyy093
Advance Access Publication Date: 13 June 2018
Software Application Profile

Software Application Profile

The *betaboost* package—a software tool for modelling bounded outcome variables in potentially high-dimensional epidemiological data

Andreas Mayr,[1]* Leonie Weinhold,[1] Benjamin Hofner,[2] Stephanie Titze,[3] Olaf Gefeller[4] and Matthias Schmid[1]

[1]Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany, [2]Section Biostatistics, Paul-Ehrlich-Institut, Langen, Germany, [3]Department of Nephrology and Hypertension and [4]Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany

*Corresponding author. Department of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität Bonn, Sigmund-Freud-Str. 25, 53105 Bonn, Germany. E-mail: amayr@uni-bonn.de

Editorial decision 25 April 2018; Accepted 11 May 2018

Abstract

Motivation: To provide an integrated software environment for model fitting and variable selection in regression models with a bounded outcome variable.

Implementation: The proposed modelling framework is implemented in the add-on package *betaboost* of the statistical software environment R.

General features: The *betaboost* methodology is based on beta-regression, which is a state-of-the-art method for modelling bounded outcome variables. By combining traditional model fitting techniques with recent advances in statistical learning and distributional regression, *betaboost* allows users to carry out data-driven variable and/or confounder selection in potentially high-dimensional epidemiological data. The software package implements a flexible routine to incorporate linear and non-linear predictor effects in both the mean and the precision parameter (relating inversely to the variance) of a beta-regression model.

Availability: The software is hosted publicly at [http://github.com/boost-R/betaboost] and has been published under General Public License (GPL) version 3 or newer.

Key words: Variable selection, model-based boosting, distributional regression, beta-regression, bounded outcome, R software

Introduction

The analysis of bounded variables such as rates or scores is a common issue in epidemiological studies. Important examples include, among many others, health-related quality of life (HRQoL) scales,[1] which are usually bounded between 0 and 100 and which will be analysed in detail in this article. Other examples are percentage measures in cancer epidemiology (such as mammographic density[2]) the Alzheimer's disease assessment scale[3] in neuroepidemiology, or DNA methylation levels ('beta values') in epigenetics.[4] Often, the goal of these studies is to relate a bounded variable to an exposure of interest, having to account for a possibly large number of potential confounders.[5] In this situation, a common approach is to fit a regression model using the bounded variable as outcome variable and the exposure, together with a subset of the confounders, as explanatory variables. The methodological challenges in this context are 2-fold: (i) the bounded outcome variable limits the use of traditional Gaussian regression techniques; and (ii) classical methods for variable selection become infeasible when there is a large number of explanatory variables. With the *betaboost* package, we present a statistical software tackling both issues by incorporating flexible approaches for beta-regression in a model-based boosting framework.[6]

For the modelling of bounded outcomes, classical Gaussian regression may lead to biased results, as the boundary constraints violate the assumption of a constant residual variance. A more flexible method that does not require variable transformations is beta-regression[7,8] which is particularly suitable for the analysis of HRQoL data and has also been used to model the other bounded outcomes mentioned above in epidemiological studies.[1,9] Beta-regression is characterized by two parameters, namely the mean parameter μ (corresponding to the expected value of the outcome) and the precision parameter φ (relating inversely to the variance, which is given by $\frac{\mu(1-\mu)}{\varphi}$). In our software, two different types of beta-regression are implemented: classical beta-regression that focuses on modelling the mean parameter, assuming the scale parameter to be constant, and an extended version that additionally relates the precision parameter to the covariates ('distributional regression'[10]). The extended version is potentially advantageous when the explanatory variables do not only affect the location of the outcome distribution but also its shape.

Common algorithms for beta-regression (as previously implemented in the *betareg*[11] package for R, the GLIMMIX procedure in SAS® or *betareg* for Stata®) are based on maximum likelihood estimation via Fisher scoring. These approaches are, however, not applicable to high-dimensional data, in particular when there are more explanatory variables than observations. A solution to deal with this problem is the use of supervised machine learning techniques[12,13] which incorporate mechanisms for simultaneous model fitting and variable selection. The software presented here implements a gradient boosting algorithm[6] which is a machine learning technique that was adapted to specifically fit statistical regression models.[14] We have demonstrated the advantages of boosting algorithms in general, and for beta-regression in particular, in various research articles.[6,15] Boosted algorithms: (i) are applicable for high-dimensional data; (ii) are able to identify the most influential explanatory variables (or confounders) in a potentially large set of candidate variables; (iii) yield model fits that have the same interpretation as fits obtained from classical modelling approaches; and (iv) allow users to incorporate different types of predictor effects (e.g. linear, splines, interactions).

With the *betaboost* package, we provide a versatile tool for the analysis of bounded outcome variables via beta-regression with boosting-based model fitting and variable selection. We illustrate these properties by analysing HRQoL data collected for the German Chronic Kidney Disease Study (GCKD[16,17]). In the Supplementary materials (available as Supplementary data at *IJE* online), we present a simulation study that illustrates the variable selection properties of *betaboost* for high-dimensional data under known conditions.

Implementation

Betaboost is an add-on package for the freely available statistical software environment R. It is a command line program implementing gradient boosting algorithms for both classical and extended beta-regression. Via wrapper functions, *betaboost* provides an interface to access the well-tested technical implementations of gradient boosting[18] (R package *mboost*) and boosting distributional regression[19] (R package *gamboostLSS*), retaining the full functionality of these packages while adding the possibility to apply beta-regression. In addition, *betaboost* provides an interface that is harmonized with standard-regression R packages, avoiding excessive technical terminology and increasing user-friendliness. For example, users can specify formulas for the desired models in a standard 'R-type' fashion, incorporating P-spline effects via *s(x)*. An example call with four explanatory variables and two P-spline effects is given by

$$\mathrm{betaboost(formula} = \mathrm{y} \sim \mathrm{x1} + \mathrm{x2} + \mathrm{s(x3)} + \mathrm{s(x4)},$$
$$\mathrm{phi.formula} = \mathrm{y} \sim \mathrm{x1} + \mathrm{x2} + \mathrm{s(x3)} + \mathrm{s(x4)}, \mathrm{data} = \mathrm{data})$$

where the *formula* and *phi.formula* arguments specify the mean and precision submodels, respectively, of an extended beta-regression model.

The core of the underlying algorithms is the optimization of the log-likelihood of a beta-regression model via component-wise gradient ascent. Here 'component-wise' means that, starting with a covariate-free model, the algorithm iteratively identifies the explanatory variables, resulting in the largest relative increase of the log-likelihood. In each iteration, only the best-performing explanatory variable is added to the model, and its effect on the outcome is estimated—already included effects get updated. As the more relevant explanatory variables are likely to be included first, variable selection is carried out. The main parameter to control variable selection is the number of boosting iterations: The larger this number, the more explanatory variables will be included in the model, and vice versa. *Betaboost* implements several data-driven functionalities to optimize the number of boosting iterations via cross-validation techniques.

In the case of extended beta-regression, the algorithm additionally evaluates for each iteration whether updating the mean model or the precision model will result in a larger increase of the log-likelihood.[20] Consequently, only the better-fitting model will be updated. In order to provide a diagnostic tool for the user to decide whether classical or extended beta-regression is more appropriate for modelling the data at hand, pseudo-R^2 measures[21] have been implemented. The *betaboost* package is hosted publicly at [http://github.com/boost-R/betaboost] and is available from [http://r-project.org].

Example

Our example is based on data collected for the German Chronic Kidney Disease Study (GCKD),[16,17] which is an ongoing cohort study that enrolled $n = 5217$ patients aged 18–74 years with stage III chronic kidney disease (CKD) or overt proteinuria/albuminuria. The aim of the study is to identify risk factors of CKD progression and to investigate the risk of cardiovascular events and death in CKD patients. Among the secondary outcome variables, health-related quality of life (HRQoL) is of particular interest. HRQoL was recorded using the self-administered KDQOL-Complete^{TM} questionnaire.[22] Here we analyse the physical health of the GCKD patients at baseline, which was measured by the Physical Composite Score (PCS) of the KDQOL-Complete, and identify a set of explanatory variables showing associations with the PCS. As the PCS is bounded between 0 ('lowest possible' physical health) and 100 ('optimal' physical health), we fit both classical and extended beta-regression models using the PCS as outcome variable. The set of potential explanatory variables consists of sociodemographic variables (e.g. education, type of health insurance), clinical variables (e.g. diagnosis/type of kidney disease) and laboratory measurements obtained from blood and urine samples (e.g. concentrations of cholesterol, calcium and haemoglobin). Altogether, there are 54 explanatory variables, leading to $2^{54} > 10^{13}$ potential predictor combinations for each model. This setting, although with far more patients than variables, is already a good example where classical subset selection strategies get practically infeasible. Patients having complete baseline data ($n = 3522$, 2141 male, 1381 female) are used for analysis. The distribution of the PCS values at baseline is left-skewed and shows clear deviations from normality (Figure S1, available as Supplementary data at *IJE* online).

To evaluate the variable selection properties of *betaboost*, we used the bootstrap[23] (sampling with replacement) to generate 1000 random samples of the baseline data. On each of the 1000 bootstrap samples we fitted both a classical beta-regression model and an extended one for the patient's PCS score. All 54 explanatory variables were considered as potential predictors (continuous variables as spline effects, categorical variables as categorical effects). On each of the bootstrap samples, we selected the optimal number of boosting iterations via the pre-implemented resampling procedures. The resulting values were 36 (median) for classical beta-regression and 276 (median) for the extended beta-regression (181 updates for the mean value, 95 updates for precision parameter). A comparison with standard software for beta-regression is not feasible here, as only *betaboost* performs variable selection. For a comparison in a smaller data set, see the Supplementary material, available as Supplementary data at *IJE* online.

To compare the performance of extended and classical beta-regression, we evaluated both models via pseudo-R^2 measures on the 1000 'out-of-bag' data sets, i.e. on the sets of observations that were not contained in the respective bootstrap samples (median R^2 classical beta-regression: 0.447, extended beta-regression: 0.455). Our approach also led to higher R^2 measures than Gaussian regression with transformed outcome variable (log transformation: 0.435, arcsine: 0.449, logit: 0.446). The estimated model size for classical beta-regression was 23 explanatory variables (median, range 11–49). In case of extended beta-regression, the estimated size of the submodel for the mean value was 26 explanatory variables (median, range: 12–46), and the estimated size of the submodel for the precision parameter was 13 explanatory variables (median, range: 4–36). According to Figure 1, high selection rates for the mean parameter were obtained by well-established predictors of HRQoL such as age, body mass index (BMI) and exercise. Also, variables associated with pain (arthritis, muscle pain) and indicators of kidney failure and

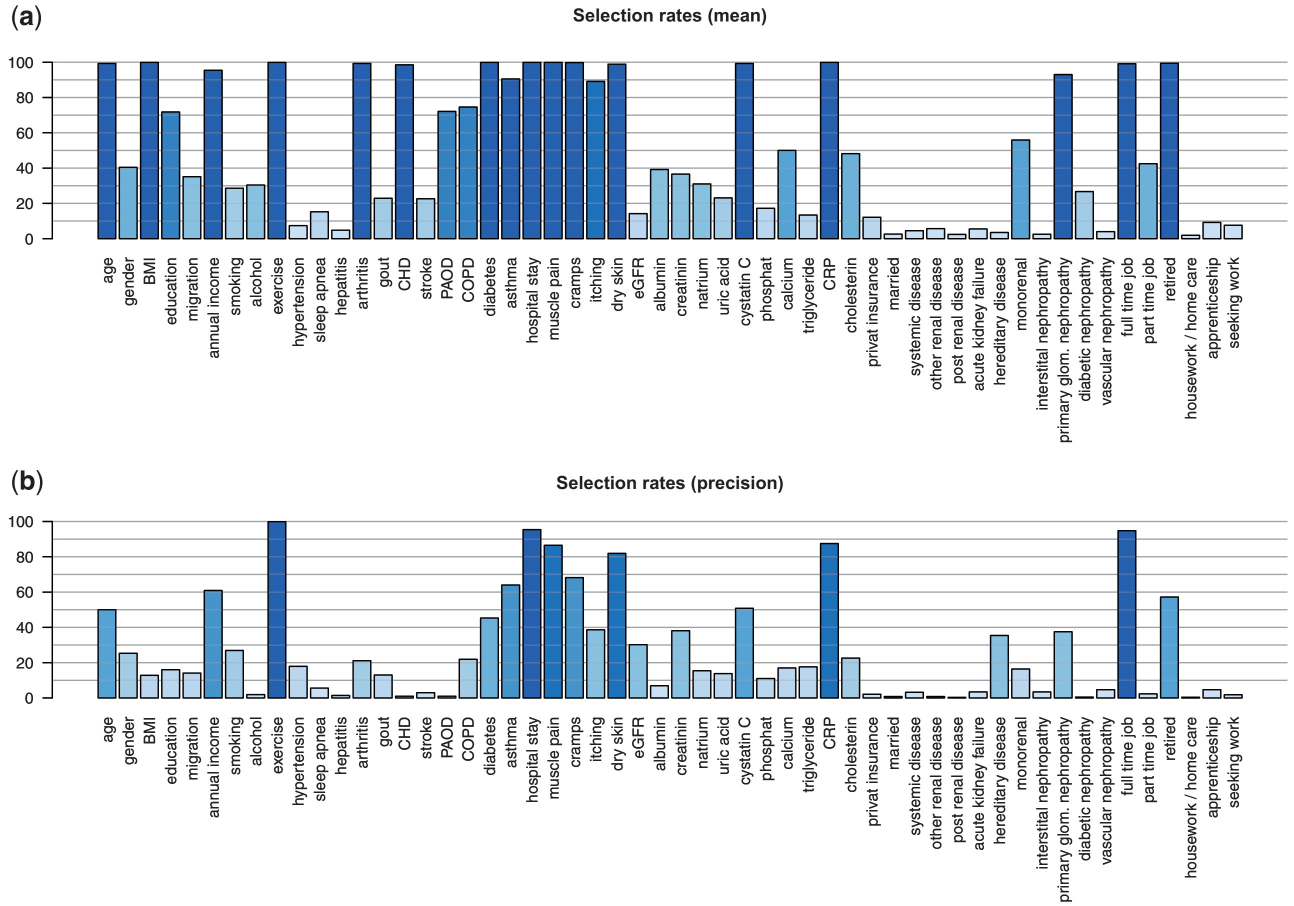

Figure 1. Analysis of the GCKD Study data. The barplots show the selection rates of the explanatory variables, as obtained from fitting extended beta regression models to 1000 bootstrap samples of the GCKD baseline data. The PCS score was used as outcome variable. The upper plot refers to the model for the mean parameter, the lower plot to the precision model.

inflammation (cystatin C, C-reactive protein) were among the variables with the highest selection rates. For the precision parameter, exercise, hospital stay and employment in a full-time job had high selection rates, indicating that these variables affect the variance of HRQoL.

As the extended beta-regression models showed better performance on the test samples (regarding the pseudo-R^2 measure), we finally fitted an extended beta-regression model on the full data set, again optimizing the number of boosting iterations. The resulting model contained 33 explanatory variables for the expected value and 21 for the precision parameter. The corresponding pseudo-R^2 was 0.505. Figure 2 displays one of the resulting non-linear spline effects, namely the one for the calcium concentration (confidence intervals are based on bootstrap samples[24]). The effect estimate shows that normal ranges of calcium concentration are associated with higher estimates of the PCS score (positive effect on location parameter), whereas hypercalcaemia as well as hypocalcaemia appear to have a negative influence on the PCS score.[25] Note that the interpretation of the effects on the precision parameter is different: the precision is inversely related to the variance of the outcome, a positive effect in this model leads to smaller variance (higher precision), whereas a negative effect leads to a larger variance (smaller precision). Other effect estimates confirm known and expected associations. For example, physical exercise has a positive effect on the PCS score, but an increasing BMI has a negative effect (Figures S2 and S3, available as Supplementary data at *IJE* online).

Discussion

Beta-regression is a state-of-the-art tool for the analysis of bounded outcome variables in epidemiological studies.[1,4] The *Betaboost* software presented in this article implements a comprehensive framework for beta-regression modelling, taking advantage of the high flexibility of gradient-boosting algorithms to allow for variable and/or confounder selection in higher-dimensional epidemiological data. *Betaboost* provides all relevant functionalities needed to run boosting algorithms for classical and extended beta-regression in one function, while sparing the researcher unnecessary technical

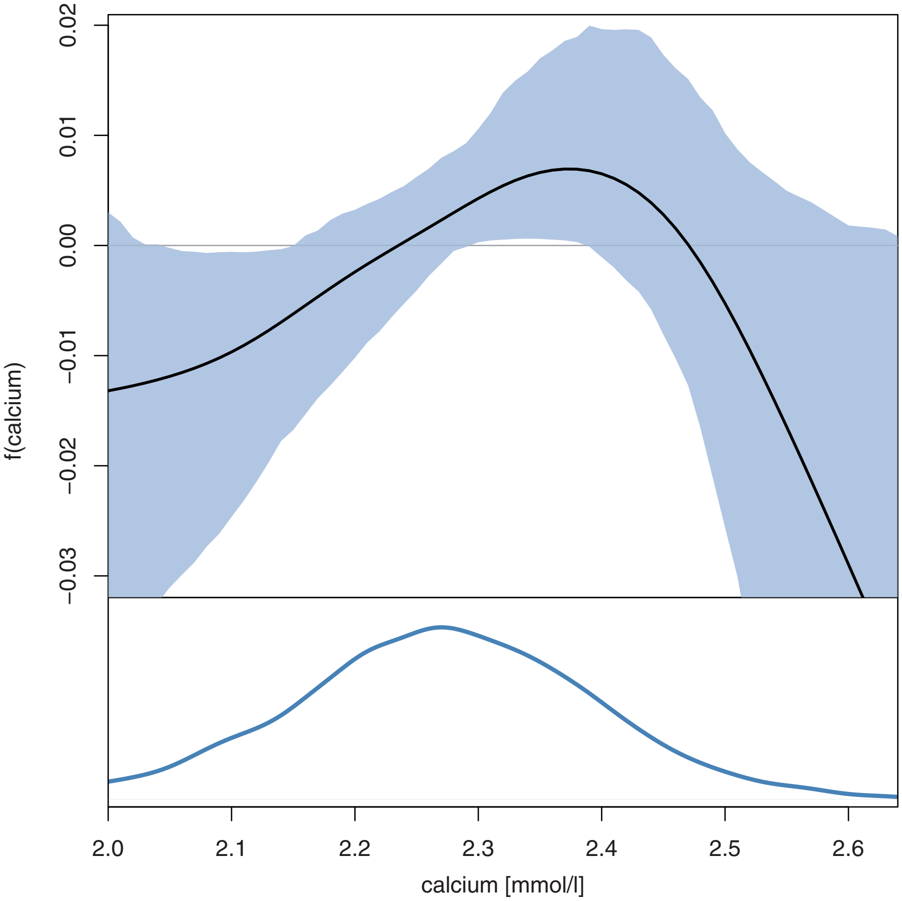

Figure 2. Analysis of the GCKD baseline data. The figure shows the P-spline estimate of the calcium concentration on the expected value of the PCS score in the extended beta regression model (logit link). The solid black line displays the effect on the complete data set. The shaded area reflects the confidence intervals estimated from the bootstrap samples. The lower plot displays the empirical distribution of the calcium concentration in the GCKD baseline data.

details. In contrast to other advanced software for beta-regression (e.g. the *betareg* package for R or its commercial competitors) *betaboost* is able to model non-linear associations between the explanatory variables and the outcome (as shown, for example, in Figure 2).

In an earlier work,[26] we have demonstrated the similarity of boosting models to penalized regression approaches such as the lasso.[27] Recently, new approaches to incorporate lasso-type penalties in extended beta-regression models have been proposed.[29] Although these methods allow for variable selection, they are restricted to linear predictor effects and are thus less flexible than *betaboost*. One limitation of the boosting approach is that, due to the iterative fitting with intrinsic variable selection, confidence intervals for effect estimates can only be computed via resampling procedures[24] (as in Figure 2). Hypothesis tests can be based on permutation procedures.[28] Another practical limitation is the considerable run-time of the algorithm for extended beta-regression, particularly for large numbers of potential predictors (see simulation study in the Supplementary material, available as Supplementary data at *IJE* online). However, to the best of our knowledge, boosting algorithms are still the only available fitting routine for beta-regression that allows for nonlinear predictor effects in the presence of high-dimensional data.

The analysis of the GCKD study data has demonstrated that *betaboost* is able to identify highly plausible associations between physical health and the characteristics of patients with medium-stage chronic kidney disease. The model fits obtained from *betaboost* are easily accessible and allow for the same interpretation as comparable models fitted via classical fitting methods.

Supplementary Material

Supplementary figures on the application example (GCKD data) and exemplary code on how to apply *betaboost*: the code reproduces the analysis of a publicly available HRQoL data set and provides a comparison with the fit of the *betareg* package on a standard data set for beta-regression.

These example analyses are also available via a vignette accompanying the package. Furthermore, we present the results of a simulation study illustrating the ability of *betaboost* to identify a small amount of truly informative predictors in a much larger set of candidate variables (pdf file). The corresponding code for the actual data analysis (GCKD data) and for the reproduction of the simulation study is also available (R files).

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

The work on this article was supported by the Deutsche Forschungsgemeinschaft [http://www.dfg.de, grant number SCHM 2966/1-2] and the Interdisciplinary Center for Clinical Research of the Friedrich-Alexander University Erlangen-Nürnberg (project J49). The GCKD study was funded by grants from the German Ministry of Education and Research (BMBF) (http://www.gesundheitsforschung-bmbf.de/de/2101.php; grant number 01ER0804) and the KfH Foundation for Preventive Medicine (http://www.kfh-stiftung-praeventivmedizin.de/content/stiftung).

Conflict of interest: None declared.

References

1. Hunger M, Doring A, Holle R. Longitudinal beta-regression models for analyzing health-related quality of life scores over time. *BMC Med Res Methodol* 2012;**12**:144.
2. Peplonska B, Bukowska A, Sobala W *et al*. Rotating night shift work and mammographic density. *Cancer Epidem Biomarkers Prev* 2012;**21**:1028–37.
3. Rogers JA, Polhamus D, Gillespie WR *et al*. Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta-regression meta-analysis. *J Pharmacokinet Pharmacodyn* 2012;**39**:479–98.
4. Campanella G, Polidoro S, Di Gaetano C *et al*. Epigenetic signatures of internal migration in Italy. *Int J Epidemiol* 2015;**44**:1442–49.
5. Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: how whole-genome sequencing and big data are transforming epidemiology. *Trends Microbiol* 2014;**22**:282–91.
6. Schmid M, Wickler F, Maloney KO, Mitchell R, Fenske N, Mayr A. Boosted Beta-regression. *PLos One* 2013;**8**:e61623.
7. Ferrari SLP, Cribari-Neto F. Beta-regression for modelling rates and proportions. *J Appl Stat* 2004;**31**:799–815.
8. Grün B, Kosmidis I, Zeileis A. Extended beta-regression in R: shaken, stirred, mixed, and partitioned. *J Stat Softw* 2012;**48**:1–25.
9. Hunger M, Baumert J, Holle R. Analysis of SF-6D index data: is beta-regression appropriate? *Value Health* 2011;**14**:759–67.
10. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc C Appl Stat* 2005;**54**:507–44.
11. Cribari-Neto F, Zeileis A. Beta-regression in R. *J Stat Softw* 2010;**34**:1–24.
12. Zhao Y, Chen F, Zhai RH *et al*. Correction for population stratification in random forest analysis. *Int J Epidemiol* 2012;**41**:1798–806.
13. Dietrich S, Floegel A, Troll M *et al*. Random survival forest in practice: a method for modelling complex metabolomics data in time to event analysis. *Int J Epidemiol* 2016;**45**:1406–20.
14. Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms from machine learning to statistical modelling. *Methods Inf Med* 2014;**53**:419–27.
15. Mayr A, Fenske N, Hofner B, Kneib T, Schmid M. Generalized additive models for location, scale and shape for high dimensional dataua flexible approach based on boosting. *J R Stat Soc C Appl Stat* 2012;**61**:403–27.
16. Eckardt KU, Barthlein B, Baid-Agrawal S *et al*. The German Chronic Kidney Disease (GCKD) study: design and methods. *Nephrol Dial Transpl* 2012;**27**:1454–60.
17. Titze S, Schmid M, Kottgen A *et al*. Disease burden and risk profile in referred patients with moderate chronic kidney disease: composition of the German Chronic Kidney Disease (GCKD) cohort. *Nephrol Dial Transpl* 2015;**30**:441–51.
18. Hofner B, Mayr A, Robinzonov N, Schmid M. Model-based boosting in R: a hands-on tutorial using the R package mboost. *Comput Stat* 2014;**29**:3–35.
19. Hofner B, Mayr A, Schmid M. gamboostLSS: an R package for model building and variable selection in the GAMLSS framework. *J Stat Softw* 2016;**74**:1–31.
20. Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B. Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *J Stat Comput* 2018;**28**:673–87.
21. Veall MR, Zimmermann KF. Pseudo-R2 measures for some common limited dependent variable models. *J Econ Surv* 1996;**10**:241–59.
22. Hays RD, Kallich JD, Mapes DL, Coons SJ, Carter WB. Development of the kidney disease quality of life (KDQOL) instrument. *Qual Life Res* 1994;**3**:329–38.
23. Efron B. 1977 Rietz lecture—bootstrap methods—another look at the jackknife. *Ann Statist* 1979;**7**:1–26.
24. Hofner B, Kneib T, Hothorn T. A unified framework of constrained regression. *Stat Comput* 2016;**26**:1–14.
25. Graciolli FG, Neves KR, Barreto F *et al*. The complexity of chronic kidney disease-mineral and bone disorder across stages of chronic kidney disease. *Kidney Int* 2017;**91**:1436–46.
26. Hepp T, Schmid M, Gefeller O, Waldmann E, Mayr A. Approaches to regularized regression—a comparison between gradient boosting and the Lasso. *Methods Inf Med* 2016;**55**:422–30.
27. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B Stat Methodol* 1996;**58**:267–88.
28. Mayr A, Schmid M, Pfahlberg A, Uter W, Gefeller O. A permutation test to analyse systematic bias and random measurement errors of medical devices via boosting location and scale models. *Stat Methods Med Res* 2015;**26**:1443–60.
29. Zhao W, Zhang R, Lv Y, Liu J. Variable selection for varying dispersion beta regression model. *J Appl Stat* 2014;**41**:95–108.