Tjeerd van der Ploeg

# Prediction of Medical Outcomes with Modern Modelling Techniques

# Prediction of Medical Outcomes with Modern Modelling Techniques

Tjeerd van der Ploeg

Lay-out and printing: Optima Grafische Communicatie, Rotterdam

# Prediction of Medical Outcomes with Modern Modelling Techniques

Het voorspellen van medische uitkomsten met moderne modelleringstechnieken

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

woensdag 11 januari 2017 om 13.30 uur

door

Tjeerd van der Ploeg
geboren te Amsterdam

**Erasmus University Rotterdam**

Erasmus

## PROMOTIECOMMISSIE

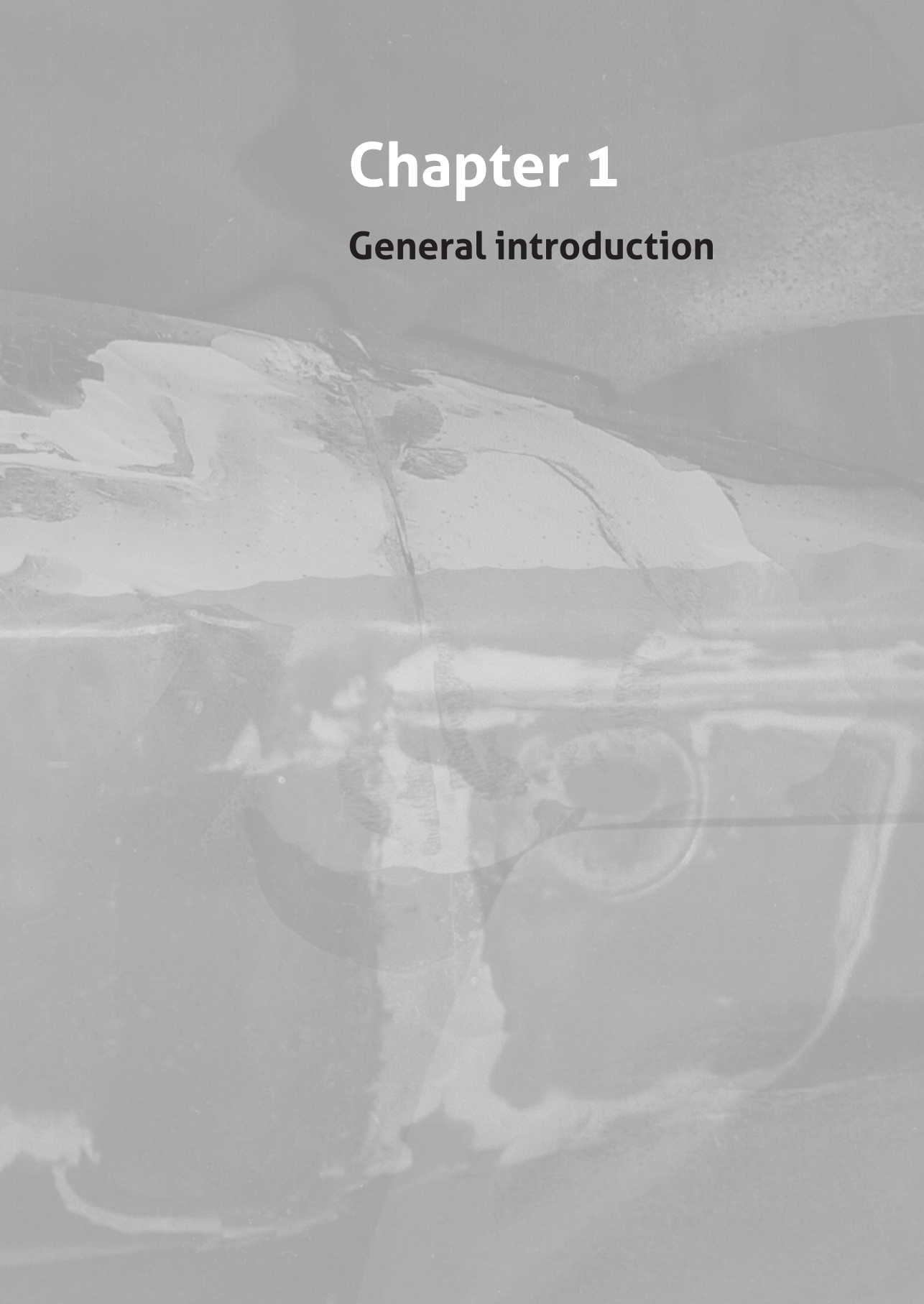
**Promotor**
Prof.dr. E.W. Steyerberg

**Overige leden**
Dr.ir. J.A. Kors
Dr. D. Rizopoulos
Prof.dr. B. van Calster

## CONTENTS

# Chapter 1

## General introduction



## GENERAL INTRODUCTION

In biomedical science and clinical practice, there is an increasing demand for accurate prediction models [1]. Prediction models are based on trends and patterns in available data to predict medical outcomes, such as complications or mortality. Prediction models may also support diagnostic classification and identify risk factors to facilitate prevention by classifying sources of infections and contaminations. Prediction models can assist physicians in making decisions by complementing their own clinical judgment with evidence-based analyses.

A large variety of modelling techniques is available nowadays, including data mining and machine learning techniques. Modellers require guidelines for selecting the appropriate tools for constructing and validating reliable prediction models. This thesis aims to give medical researchers insight into the pros and cons of modern modelling techniques versus traditional modelling techniques [2] [3] [4]. These insights may help to determine whether a published decision tool is valid and to advise researchers on the role of modern modelling techniques in various settings, such as a limited or larger sample size [5] [6] [7].

### Traditional prediction modelling

In biomedical science, traditional and frequently used modelling techniques are linear regression, logistic regression and Cox regression. These regression techniques are all based on a linear combination of the predictor variables, the so called linear predictor. For $p$ independent predictor variables $x_1,...,x_p$, the linear predictor ($lp$) takes the form: $lp = b_1 * x_1 + ... + b_p * x_p$, in which $b_1,...,b_p$ are the regression coefficients for the $p$ predictor variables.

A linear regression model can be written as:
$y = b_0 + lp + \varepsilon$, in which $\varepsilon$ is the error variable and $b_0$ refers to the intercept, also indicated with $\alpha$ sometimes. The coefficients $b_0,...,b_p$ are calculated by minimizing $\Sigma\varepsilon^2$. The dependent variable $y$ is continuous and the independent variables are continuous, categorical or dichotomous.

A logistic regression model can be written as:
$P(y=1) = \frac{1}{1+e^{-(b_0+lp)}}$, in which $P(y=1)$ is the probability that y=1.
The dependent variable $y$ is dichotomous (0/1) and the independent variables are continuous, categorical or dichotomous. The coefficients $b_0,...,b_p$ are calculated by maximizing the likelihood.

A Cox regression model is the most often used method for survival outcomes and can be written as: $H(t) = H_0(t) * e^{lp}$, in which $H(t)$ is the hazard at time $t$ and $H_0(t)$ is the baseline hazard function at time $t$. The independent variables are continuous, categorical or dichotomous. The coefficients $b_1,...,b_p$ are calculated by maximizing the likelihood.

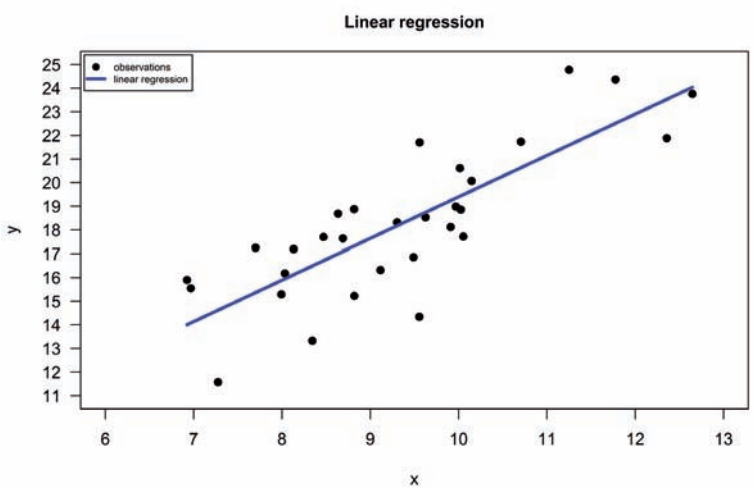



**Figure 1** Example of a linear regression analysis in 30 patients

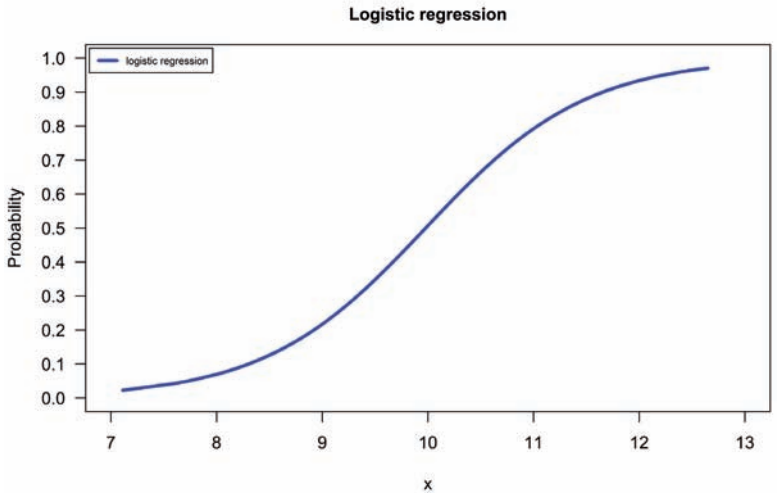



**Fig 2** Example logistic regression

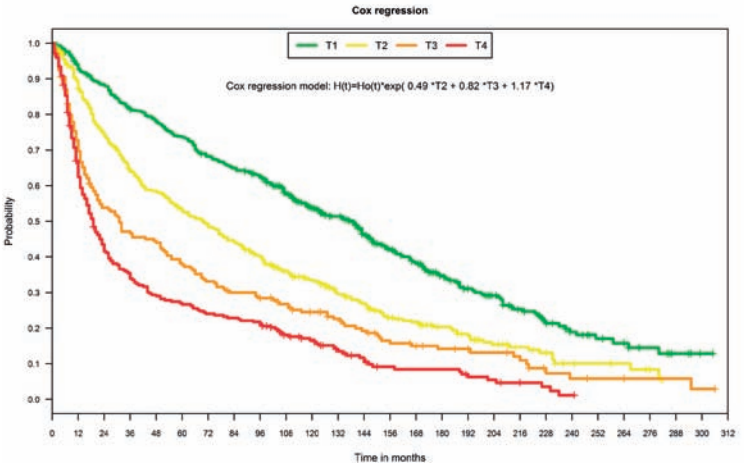



**Fig 3** Example Cox regression with four subgroups

A less frequently used traditional modelling technique is classification and regression trees (CART) [8]. CART is a modelling technique that uses recursive partitioning to split the patient records that serve as a training data set into segments with similar endpoint values. The modelling starts by examining the input variables to find the best split, commonly measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two further subgroups and so on, until a stopping criterion is met. The dependent variable can be continuous or categorical. The independent variables are continuous, categorical or dichotomous, but they are always dichotomized in the analysis. Figure 4 shows an example of a tree model with three predictors (A, B and C) and one outcome, all with categories a, b, and c.

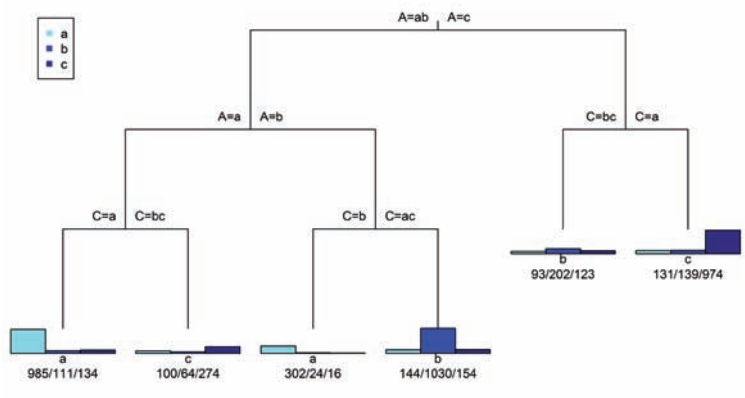



**Fig 4** Example tree model CART

## Modern prediction modelling

Predictive data mining has received increasing interest as an instrument for researchers across various fields. Nowadays, there is a widespread availability of new computational methods and tools for data analysis and predictive modelling. In particular, methods known as "data mining" or "machine learning" offer new methodological and technical solutions for the analysis of medical data and the construction of prediction models. Examples of these techniques include random forests, support vector machines and neural networks. These techniques are based on algorithms which operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions as used in traditional regression modelling. Some general descriptions are given below. Further details are presented in the chapters of this thesis.

### *Random forests*

The random forest technique (RF) is an ensemble classifier that consists of many classification trees, rather than a single tree such as CART. Each tree is constructed using a bootstrap sample from the original data (a sample drawn with replacement). For classification, random forest outputs the class that is the mode among the classes from individual trees. In case of a regression-based prediction, a random forest outputs the value that is the mean of the output values from individual trees [9].

For each tree the misclassification rate is calculated using the subjects that are not in the bootstrap sample, approximately 36.8% of the original data. The misclassification rate is called the "out of bag" error rate (OOB). The overall OOB error rate is calculated by aggregation over the trees. Key parameters for the random forest technique are the number of trees and the number of candidate variables. The default setting for the number of candidate variables is the square root of the total number of all predictors in case of classification or the total number of all predictors divided by 3 in case of regression. Random forests can also be used to rank the importance of the predictor variables by means of a variable importance plot. The importance of a predictor variable is calculated by the mean decrease in accuracy or the mean decrease in Gini of the model. The mean decrease in accuracy or Gini represents how much the accuracy or Gini of the model is reduced by removing the variable. The dependent variable and the independent variables are continuous or categorical.

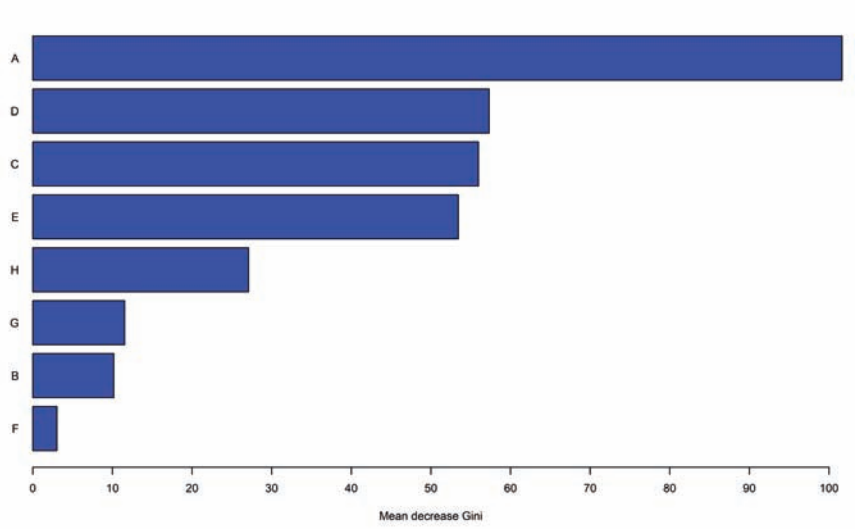



**Fig 5** Example variable importance plot with 8 predictor variables

### *Support vector machines*

A support vector machine (SVM) performs classification tasks by constructing hyperplanes with a margin in a multidimensional space which separate cases from different classes. An SVM can perform a non-linear classification or regression task using different kernels (radial, linear and polynomial). The tuning parameters for SVMs are the C-parameter (cost), which regulates the margin width, and the gamma-parameter for the kernel calculation. SVM claims to be a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM may be particularly suited for analysing data with large numbers of predictor variables. The dependent variable and the independent variables are continuous or categorical. SVM uses the distance from each data point to the decision boundary to calculate the so-called decision value. With this decision value a prediction is made (positive decision value: outcome=1 and negative decision value: outcome=-1). These decision values can also be used to calculate probabilities for the outcome category of interest (-1 of 1) [10].

Figure 6 shows examples of a SVM with a linear kernel, two predictor variables ($x_1$ and $x_2$) and a dichotomous outcome with different settings for the cost parameter (gamma parameter=0.5 for a linear SVM). The figure shows that a lower cost-value leads to a wider margin and therefore to more misclassification [11].

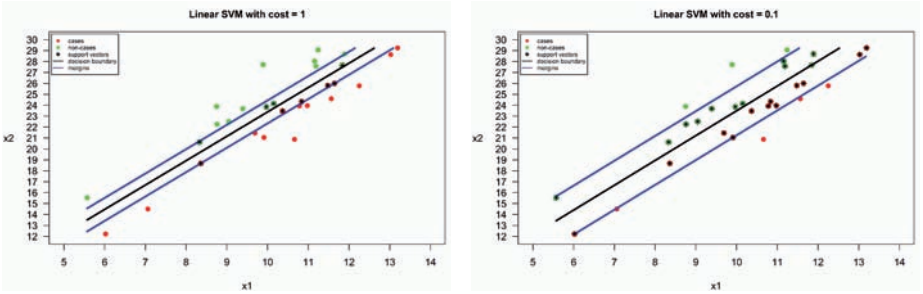



**Figure 6** Examples of a linear SVM

Figure 7 shows examples of a SVM with a radial based function kernel, two predictor variables ($x_1$ and $x_2$) and a dichotomous outcome with different settings for the cost parameter and the gamma parameter. The figure shows that a higher value for the cost parameter and a lower value for the gamma parameter lead to a wider margin and therefore to more misclassification.

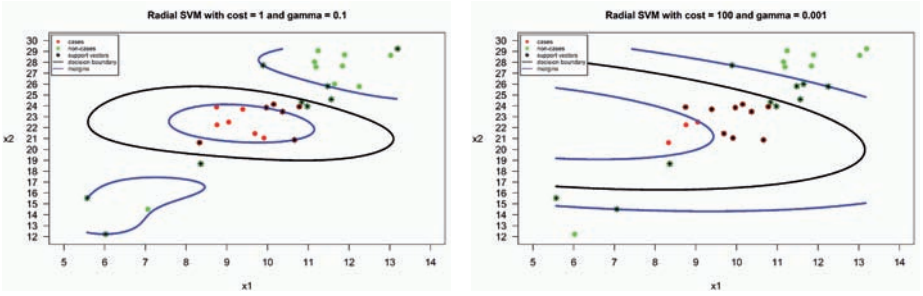



**Figure 7** Examples of a radial SVM

*Neural networks*

A neural network (NN), sometimes called a multilayer perceptron, simulates a large number of interconnected simple processing units, which are arranged in layers. There are three parts in a neural network: an input layer, with units representing the predictor variables, one or more hidden layers, and an output layer, with a unit representing the endpoint. The units are connected with varying connection strengths or weights. Input data are presented to the input layer and values are propagated from there to the next layer. Then, a prediction is delivered from the output layer. The network learns by examining individual records, generating a prediction for each record and making adjustments to the weights whenever an incorrect prediction is made. The adjustments are commonly based on the gradient descent algorithm to minimize the prediction error. This process is repeated many times, and the network continues to improve its

predictions until the magnitude of the gradient is less than a chosen bound (0.00005 e.g.). The crucial parameters of an NN are the size parameter (number of units in the layer) and the decay parameter that penalizes large weights in the model to avoid overfitting (default=0). The dependent variable and the independent variables can be continuous, categorical or dichotomous [12].
Figure 8 shows an example of a neural network with an input layer consisting of two input units (green), a hidden layer with three units (blue) and an output layer with one output unit (red). The input variables are $x_1$ and $x_2$ and the output variable is y.

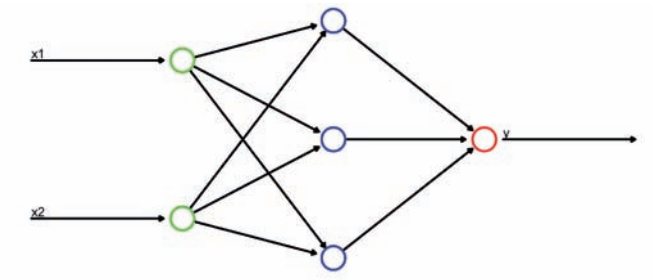



**Figure 8** Example of a neural network

A value $y_j$ for a unit in the hidden layer (blue) is calculated using a linear combination of the values $x_1$ and $x_2$ of the units in the input layer (green) with coefficients (weights) $w_{0j}$, $w_{1j}$ and $w_{2j}$.
In formula: $y_j = w_{0j} + w_{1j} * x_1 + w_{2j} * x_2, \quad j=1,2,3$

The value z is then calculated using a linear combination of the values $y_1$, $y_2$ and $y_3$ of the units in the hidden layer (blue) with coefficients $w_{30}$, $w_{31}$, $w_{32}$ and $w_{33}$.
In formula: $z = w_{30} + w_{31} * y_1 + w_{32} * y_2 + w_{33} * y_3$

The value of z is then compared with the value of y in the unit of the output layer by means of the prediction error defined as $\frac{1}{2}*(y-z)^2$.

The best model is found by minimizing this prediction error by means of repeated adjustment of the weights $w_{ij}$ in the negative direction of the gradient until the magnitude of the gradient is less than a chosen bound (0.00001 e.g.). In case of a dichotomous outcome variable, the formulas for the calculation of $y_j$ and z become:

$$y_j = \frac{1}{1+e^{-(w_{0j} + w_{1j} * x_1 + w_{2j} * x_2)}} \quad , j = 1,2,3 \quad \text{and} \quad z = \frac{1}{1+e^{-(w_{30} + w_{31} * y_1 + w_{32} * y_2 + w_{33} * y_3)}}$$

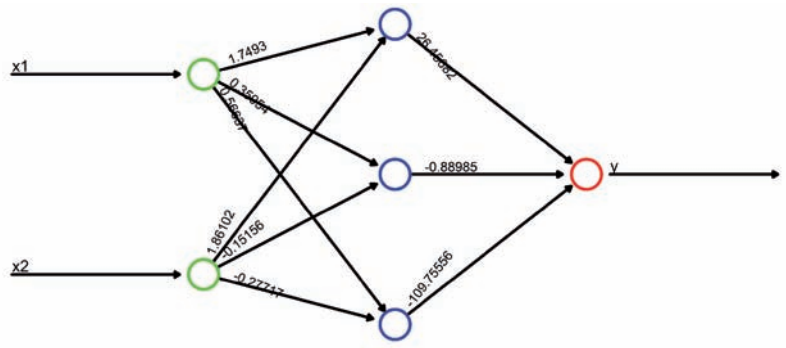



**Figure 9** Neural network with weights

Figure 9 shows an example of a neural network with the calculated weights.

*Least absolute shrinkage selection operator*

The least absolute shrinkage selection operator (LASSO) is a variant of linear or logistic regression technique. For *p* independent predictor variables $x_1, x_2, ..., x_p$ and an outcome variable y, LASSO fits a linear model:

$y = b_0 + b_1 * x_1 + ... + b_p * x_p + \varepsilon$, in which $\varepsilon$ is the error variable and $b_0$ refers to the intercept. The regression coefficients $b_0,...,b_p$ are calculated by minimizing $\sum \varepsilon^2$ subject to

$$\sum_{j=1}^{p} |b_j| <= s .$$

In the constraint, *s* is a chosen bound (0.01 e.g.). A small value for *s* leads to small values of the regression coefficients. Often, some of the coefficients $b_j$ are shrunk to zero. The dependent variable y is continuous and the independent variables are continuous or dichotomous. Cross-validation is used to estimate the best value for *s*. In case of a dichotomous outcome, LASSO fits a logistic regression model:

$$P(y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 * x_1 + ... + b_p * x_p)}}$$ by maximizing the likelihood subject to $\sum_{j=1}^{p} |b_j| <= s$

*Bayes network*

A Bayes network is a graphical model that displays the relation of predictor variables and outcome variables and the probabilistic dependencies between these variables. A Bayes network may represent causal relationships between the variables. The links,

however, do not necessarily represent direct cause and effect. Figure 10 shows an example of a Bayes network with the probabilistic dependencies between symptoms and disease which can be used to calculate the probability of a patient having a specific

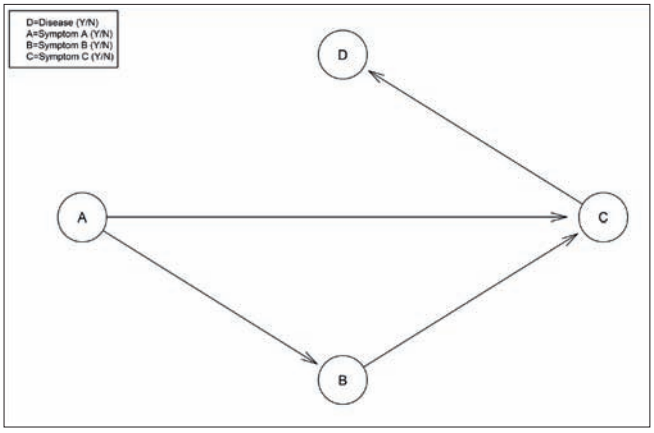



**Fig 10** Example of a Bayes network

disease (D), given the presence (Y/N) of certain symptoms (A, B and C). These probabilities are usually shown in the conditional probability tables presented with the graph. The joint probability, using Bayes theorem for conditional probabilities, can be written as:

$$P(A,B,C,D) = P(D|A,B,C) * P(C|A,B) * P(B|A) * P(A)$$

For this network, this formula reduces to:

$$P(A,B,C,D) = P(D|C) * P(C|A,B) * P(B|A) * P(A)$$

because there is only a probabilistic dependency between D and C.
With this formula, conditional probabilities such as $P(D=Y|B=Y)$ can be calculated:

$$P(D=Y|B=Y) = \frac{P(B=Y,D=Y)}{P(B=Y)} = \frac{\sum_{\{A,C\}} P(A,B=Y,C,D=Y)}{\sum_{\{A,C,D\}} P(A,B=Y,C,D)}$$

where A, C, and D are varied over the categories (Y/N).

### *Decision list*

A decision list technique (DL) outputs a set of conjunctive rules for classification. Each rule splits the data into subgroups that show a higher or lower likelihood of a dichotomous outcome relative to the overall sample. A rule may consist of more than one condition. The rules must be applied in the order listed by the model to avoid ambiguity. The outcome for a new unseen patient is determined by the first matching rule. If no matching rule can be found, the case is assigned to the so-called remainder rule. The methods used to generate the rule sets and the measures used in ordering the rules may be different for different decision list techniques. The basic principle of these techniques is the same: recognizing characteristics based on training data.
Example of a decision list output:
(Condition 1) and (Condition 2) -> (Class=C1)
(Condition 3) -> (Class=C2)
(Condition 1) -> (Class=C1)

### *Chi-squared automatic interaction detection*

The Chi-squared automatic interaction detection model (CHAID) is a classification method for building decision trees based on Chi-square analysis. CHAID first performs a univariate analysis for each of the predictor variables with respect to the outcome variable. If more than one of these relations is statistically significant (p-value<0.05 e.g.), CHAID will select the predictor with the smallest p-value. If a predictor has more than two categories, these are compared, and categories that show a similar outcome are combined. This is done by successively combining the pair of categories showing the least significant difference. CHAID is applicable for categorical and continuous predictors.
The main difference between CHAID and CART is in the construction of decision tree. CHAID uses a stopping rule (p-value<0.05 e.g.), while CART first grows the full tree and then prunes the tree based on the comparison of the performance of the tree on the development set and the performance on a holdout set. The pruning stops when equal performance is achieved.

**Table 1** Characteristics modelling techniques

| | | Modelling technique | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Linear regression | Logistic regression | Cox regression | CART | Random forests | Support vector machine | Neural net | LASSO | Bayes network | Decision list | CHAID |
| Categorizing continuous predictor variables | | - | - | - | + | + | - | - | - | + | + | + |
| Outcome | Continuous | + | - | - | + | + | + | + | + | + | + | + |
| | Categorical | - | + | - | + | + | + | + | + | + | + | + |
| | Dichotomous | - | + | - | + | + | + | + | + | + | + | + |
| | Time to event | - | - | + | - | - | - | - | - | - | - | - |
| Iteractions | Assumed | - | - | - | + | + | - | + | - | + | + | + |
| | Flexible | + | + | + | - | - | + | - | + | - | - | - |
| Selection of predictor variables | Assumed | - | - | - | + | + | - | - | - | + | + | + |
| | Flexible | + | + | + | - | - | + | + | + | - | - | - |
| Formula | | + | + | + | - | - | - | - | + | - | - | - |

+=yes, -=no

**Table 2** Hyper parameters modelling techniques

| Modelling technique | Hyper parameters |
|---|---|
| Linear regression | L2-penalization |
| Logistic regression | L2 penalization |
| Cox regression | L2-penalization |
| CART | complexity, tree depth, parent and child node size |
| Random forests | number of trees, number of variables tried |
| Support vector machines | margin width (C-parameter), gamma, kernel type |
| Neural net | number of layers, layer size, decay, learning rate |
| LASSO | upper bound L1-penalization |
| Bayes network | - |
| Decision list | - |
| CHAID | tree depth, parent and child node size, p-value for split |

In 2012, a systematic review by Bouwmeester et al. [13] revealed that most of the prediction models used in clinical prediction research were developed with traditional modelling techniques, such as linear regression, logistic regression or Cox regression. However, another review illustrated a growing trend to apply machine learning techniques in cancer research, where techniques used for feature selection and classification included neural networks, Bayesian networks, support vector machines and decision trees [14].

### Prediction of survival

Table 1 shows that modern modelling techniques are not suited for time-to-event outcomes. The use of modern modelling techniques in survival problems is complicated by the fact that these models require a single outcome variable, whereas the outcome of survival problems involves a time-to-event variable. This problem can be solved by transforming the time-to-event outcome into new single variables, so-called pseudo values [15]. These pseudo values can be used to develop machine learning models for survival problems.

### Feature selection

The selection of important predictor variables or features is sometimes embedded in the modelling technique (Table 1). A special case in developing prediction models is feature selection in a setting in which the number of predictors ($p$) is higher than the number of subjects ($n$), the "p>n" problem. A common approach is preselecting relevant features using a univariate technique with respect to the outcome (T-test, Mann-Whitney-test, Pearson correlation coefficients). By contrast, a specific modelling technique can be used to select features, and with the selected feature set a model can be built with that same modelling technique. Popular feature selection methods nowadays are the "least absolute shrinkage and selection operator" method (LASSO), recursive feature elimination with support vector machines (SVM RFE), and a backward feature selection method based on random forests (VARSEL RF) [16] [17] [18]. Their relative performance is insufficiently known, as is the performance of alternative approaches.

### Performance measures for prediction models

Since prediction models are intended to be used as decision tools for clinical practice, clinicians have to be able to determine the quality of a published decision tool. For assessing the quality of prediction models, many performance measures have been described [19]. The performance of prediction models for binary outcomes is commonly measured with discriminatory ability (AUC), the Brier score or Nagelkerke's $R^2$. The

performance of prediction models for continuous outcomes is commonly measured with the mean squared error (MSE) or the $R^2$ statistic.
Another important aspect in assessing the quality of predictive modelling is calibration. For a well-calibrated model, the predicted probabilities must match the actual probabilities: do close to *x* of 100 patients with a risk prediction of *x* % have the outcome of interest? The Cox recalibration framework is useful for assessing the calibration of a model [20], in combination with graphical assessments [21] [22].

### Sample size at model development

An important aspect in developing prediction models is the required sample size of the development set. For logistic regression modelling, an often-used rule of thumb is 10 events per variable (EPV) [23] [24]. For other modelling techniques, such a rule of thumb is not available. The higher flexibility of modern modelling techniques implies that larger sample sizes may be required for reliable estimation. To determine whether a given data set is sufficient for developing a prediction model with a good and stable predictive performance, researchers need insight into the "data hungriness" of various modern modelling techniques.

### Validation

A prediction model is only useful if the model is able to predict the correct outcome for new, unseen patients. Therefore, the validity of a prediction model is a very important issue. Two types of validation can be distinguished: internal validation and external validation. In case of internal validation, the model performance is estimated for the patients from the underlying population involved in the development of the prediction model. Commonly used techniques for internal validation are cross-validation and bootstrap resampling. These techniques can assess and, if necessary, correct for a model's optimism. However, to determine whether a particular model is transportable to slightly different settings, internal validation is not sufficient. It needs to be supplemented by external validation, in which the model is tested on one or more external data sets [25] [26].

### Aim of this thesis

The aim of this research is to investigate in what circumstances and under what conditions relatively modern modelling techniques such as support vector machines, neural networks and random forests have advantages in medical prediction research over more classical modelling techniques, such as linear regression, logistic regression and Cox regression.

## Specific research questions:

Question 1:
Comparison of modern and traditional modelling techniques:
- What is the performance in predicting intracranial findings on CT scans?
- What is the ability to capture nonlinearity?

Question 2:
Application of modern modelling techniques:
- How can they be applied for survival problems?
- How can they be applied for feature selection in a domain with many variables and comparatively few subjects or data points?

Question 3:
Performance of modern modelling techniques:
- What is the performance in relation to the sample size?
- What is the stability of the performance at external validation?

**Case studies**

To address these research questions, we performed studies with different data sets, which are briefly described below.

*HNSCC survival*

We analyzed survival in 1282 Dutch patients with newly diagnosed Head and Neck Squamous Cell Carcinoma (HNSCC) with conventional Kaplan-Meier and Cox regression analysis and modern modelling techniques. We considered clinical predictor variables such as TNM-classification, tumor location and demographic factors for predicting 5-year mortality and overall mortality.

*Legionella strains*

We analyzed a data set containing 222 Legionella pneumophila strains with 448 continuous markers. We aimed to predict a dichotomous outcome (clinical or environmental Legionella).

*CT scanning in TBI*

We investigated whether alternative modelling techniques might improve the performance of prediction rules for intracranial traumatic findings in patients with minor head injury. We re-analyzed 3181 patients with minor head injury who had received CT scans between February 2002 and August 2004. Of these patients, 243 (7.6%) had intracranial traumatic findings and 17 (0.5%) needed a neurosurgical intervention.

*Moderate or severe TBI*

We performed simulation studies based on 1731 patients with traumatic brain injury (6-month mortality 22%). We further performed external validation studies within the IMPACT data base, which comprises data of fifteen different studies. Patients were enrolled in one of ten randomized clinical trials or in one of five registries between 1984 and 2006.

**Table 3** Summary characteristics of data sets used in this thesis

| | Data set | | | |
|---|---|---|---|---|
| | HNSCC survival | Legionella strains | CT scanning in TBI | Moderate or severe TBI |
| Size | 1282 | 222 | 3181 | 11026 |
| Outcome | Mortality | Origin strain | Intracranial findings | Mortality |
| Number of categorial predictors | 7 | 0 | 10 | 2-7 |
| Number of continuous predictors | 1 | 448 | 2 | 1-3 |
| Modelling techniques | LR | LR | LR | LR |
| | CART | CART | CART | CART |
| | RF | RF | SVM | RF |
| | SVM | SVM | BN | SVM |
| | NN | NN | NN | NN |
| | | VARSEL RF | CHAID | |
| | | SVM RFE | DL | |

LR: logistic regression
CART: classification and regression trees
SVM: support vector machines
RF: random forests
NN: neural networks
CHAID: chi square automated interaction detection
BN: Bayes network
DL: decision list
VARSEL RF: variable selection random forest
SVM RFE: support vector machine recursive feature elimination

## Thesis structure

This thesis has seven core chapters (chapters 2 to 8) that address the key research questions. Some include simulation studies, others present case studies to serve as a basis for more general conclusions, which are discussed in chapter 9.

## REFERENCES


1. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG: Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med 2013, 10.
2. Austin PC, Tu J V, Ho JE, Levy D, Lee DS: Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol 2013, 66:398–407.
3. Austin PC, Lee DS, Steyerberg EW, Tu J V.: Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? Biometrical J 2012, 54:657–673.
4. Breiman L, others: Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat Sci 2001, 16:199–231.
5. Bellazzi R, Zupan B: Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008, 77:81–97.
6. Wyatt JC, Altman DG: Commentary: Prognostic models: clinically useful or quickly forgotten? Bmj 1995, 311:1539–1541.
7. Lavrač N: Selected techniques for data mining in medicine. Artif Intell Med 1999, 16:3–23.
8. Breiman L, Friedman JH, Olshen RA, Stone CJ: Classification and Regression Trees. Volume 19; 1984.
9. Breiman LEO: Random Forests. Mach Learn 2001, 45:5–32.
10. Platt J, others: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv large margin Classif 1999, 10:61–74.
11. Cortes C, Vapnik V: Support-vector networks. Mach Learn 1995, 20:273–297.
12. Harrison RF, Kennedy RL: Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Ann Emerg Med 2005, 46:431–439.
13. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, Moons KGM: Reporting and Methods in Clinical Prediction Research: A Systematic Review. PLoS Med 2012, 9:e1001221.
14. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V, Fotiadis DI: Machine Learning Applications in Cancer Prognosis and Prediction. Comput Struct Biotechnol J 2014.
15. Andersen PK, Hansen MG, Klein JP: Regression analysis of restricted mean survival time based on pseudo-observations. In Lifetime Data Anal. Volume 10; 2004:335–350.
16. Wang HY, Zheng H, Azuaje F: Evaluation of computational classification methods for discriminating human heart failure etiology based on gene expression data. 2006 Comput Cardiol 2006.

17. Guyon I, Elisseeff A: An introduction to variable and feature selection. J Mach Learn Res 2003, 3:1157–1182.

18. Diaz-Uriarte R: GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics 2007, 8:328.

19. 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 traditional and novel measures. Epidemiology 2010, 21:128–138.

20. 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 traditional and novel measures. Epidemiology 2010, 21:128–38.

21. Harrell FE, Lee KL, Mark DB: Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996, 15:361–387.

22. Austin PC, Steyerberg EW: Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 2014:0962280214558972.

23. Harrell FE: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer; 2001.

24. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD: Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000, 19:1059–1079.

25. Steyerberg EW, Vergouwe Y: Towards better clinical prediction models: Seven steps for development and an ABCD for validation. Eur Heart J 2014:1925–1931.

26. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Ann Intern Med 1999, 130:515–524.

# Chapter 2

## Prediction of intracranial findings on CT-scans by alternative modelling techniques

Tjeerd van der Ploeg, Marion Smits, Diederik W. Dippel, Myriam Hunink, Ewout W. Steyerberg

BMC Med Res Methodol. 2011 Oct 25;11:143. doi: 10.1186/1471-2288-11-143.

## ABSTRACT

### Background

Prediction rules for intracranial traumatic findings in patients with minor head injury are designed to reduce the use of computed tomography (CT) without missing patients at risk for complications. This study investigates whether alternative modelling techniques might improve the applicability and simplicity of such prediction rules.

### Methods

We included 3181 patients with minor head injury who had received CT scans between February 2002 and August 2004. Of these patients 243 (7.6%) had intracranial traumatic findings and 17 (0.5%) underwent neurosurgical intervention. We analyzed sensitivity, specificity and area under the ROC curve (AUC-value) to compare the performance of various modelling techniques by 10x10 cross-validation. The techniques included logistic regression, Bayes network, Chi-squared Automatic Interaction Detection (CHAID), neural net, support vector machines, Classification And Regression Trees (CART) and "decision list" models.

### Results

The cross-validated performance was best for the logistic regression model (AUC 0.78), followed by the Bayes network model and the neural net model (both AUC 0.74). The other models performed poorly (AUC<0.70). The advantage of the Bayes network model was that it provided a graphical representation of the relationships between the predictors and the outcome.

### Conclusions

No alternative modelling technique outperformed the logistic regression model. However, the Bayes network model had a presentation format which provided more detailed insights into the structure of the prediction problem. The search for methods with good predictive performance and an attractive presentation format should continue.

## 2.1 BACKGROUND

Minor head injury is one of the most common injuries seen in western emergency departments. Patients with minor head injury include those with blunt injury to the head who have a normal or minimally altered level of consciousness on presentation at the emergency department. Intracranial complications after minor head injury are infrequent, but they commonly require in-hospital observation and occasionally even neurosurgical intervention.

The imaging procedure of choice for reliable, rapid diagnostics of intracranial complications is computed tomography (CT). However, it is inefficient to scan all patients with minor head injury to exclude intracranial complications, as most patients with minor head injury do not show traumatic abnormalities on CT.

Several prediction rules have been developed to identify those at risk of abnormalities on CT. These include the CT in Head Injury Patients (CHIP) prediction rule [1], the Canadian CT Head Rule (CCHR) [2] and the New Orleans Criteria (NOC) [3]. While the NOC was developed by expert opinion and based on existing literature, the CCHR and CHIP rules were developed with recursive partitioning (Classification And Regression Trees, CART) and logistic regression techniques respectively (Table1).

**Table 1** Rules

| Rule | Patient selection | N patients | N predictors considered | N predictors included | Modelling technique |
|---|---|---|---|---|---|
| NOC | Prospective cohort study | 520 | >7 | 7 | Expert opinion |
| CCHR | Prospective cohort study | 3121 | 24 | 7 | Logistic regression/CART |
| CHIP | Prospective cohort study | 3181 | 14 | 14 | Logistic regression |
| Lancet | Prospective cohort study | 42411 | 10 | 3 | CART |

A recent study used CART modelling to develop a prediction rule for CT scanning in children [4]. CART modelling was argued to be a more appropriate method for the particular problem of selecting a very low risk group among patients with possible intracranial complications.

We hypothesized that alternative modelling techniques might deliver better results in terms of applicability and performance than modelling based on conventional modelling techniques such as logistic regression techniques. We compared logistic regression modelling to alternative modelling techniques [5] [6], including CART and six other techniques, in the context of selective CT scanning for minor head injury. Data from the CHIP study, underlying the CHIP prediction rule, were used for this purpose.

**Table 2** Patient characteristics

| | | Intracranial lesions | | | | |
|---|---|---|---|---|---|---|
| | | *absent* | | *present* | | |
| | | *n* | *(%)* | *n* | *(%)* | *p-value* |
| Fracture skull | Absent | 2901 | (98.7) | 207 | (85.2) | 0.000 |
| | Present | 37 | (1.3) | 36 | (14.8) | |
| EMV presentation (total) = 13 | Absent | 2818 | (95.9) | 212 | (87.2) | 0.000 |
| | Present | 120 | (4.1) | 31 | (12.8) | |
| EMV presentation (total) = 14 | Absent | 2447 | (83.3) | 166 | (68.3) | 0.000 |
| | Present | 491 | (16.7) | 77 | (31.7) | |
| Memory deficit | Absent | 2535 | (86.3) | 171 | (70.4) | 0.000 |
| | Present | 403 | (13.7) | 72 | (29.6) | |
| Contusion skull | Absent | 1863 | (63.4) | 103 | (42.4) | 0.000 |
| | Present | 1075 | (36.6) | 140 | (57.6) | |
| Loss of consciousness | Absent | 1169 | (39.8) | 61 | (25.1) | 0.000 |
| | Present | 1769 | (60.2) | 182 | (74.9) | |
| Seizure | Absent | 2920 | (99.4) | 238 | (97.9) | 0.000 |
| | Present | 18 | (0.6) | 5 | (2.1) | |
| Vomiting | Absent | 2651 | (90.2) | 188 | (77.4) | 0.000 |
| | Present | 287 | (9.8) | 55 | (22.6) | |
| Coumarins | Absent | 2868 | (97.6) | 230 | (94.7) | 0.005 |
| | Present | 70 | (2.4) | 13 | (5.3) | |
| Neurological deficit (all) | Absent | 2676 | (91.1) | 201 | (82.7) | 0.000 |
| | Present | 262 | (8.9) | 42 | (17.3) | |
| Cause | Reference | 1882 | (64.1) | 102 | (42) | 0.000 |
| | ped.or cyclist | 297 | (10.1) | 51 | (21) | |
| | Fall | 702 | (23.9) | 82 | (33.7) | |
| | Ejected | 57 | (1.9) | 8 | (3.3) | |
| PTA in 3 categories | <= 2 hrs | 2910 | (99.0) | 232 | (95.5) | 0.000 |
| | >2 hrs and <= 4 hrs | 25 | (0.9) | 6 | (2.5) | |
| | >4 hrs | 3 | (0.1) | 5 | (2.1) | |
| | | *mean* | *(sd)* | *mean* | *(sd)* | *p-value* |
| EMV change | | 0.07 | (0.50) | −0.04 | (1.27) | 0.186 |
| Age - 16 per decade | | 2.48 | (1.85) | 3.22 | (2.01) | 0.000 |

## 2.2 METHODS

The CHIP database contains data on 3181 patients with minor head injury, defined as a presenting Glasgow Coma Scale (GCS) score of 13 to 15, and a maximum loss of consciousness of 15 minutes, posttraumatic amnesia for 60 minutes. Several risk factors were recorded to predict the presence of intracranial traumatic findings on CT (Table 2). Most of the risk factors were dichotomous variables (absent, present) and a few were continuous. The outcome of interest was intracranial traumatic findings on CT (absent, present). These intracranial traumatic findings included contusions, small haemorrhages indicating diffuse axonal injury, subarachnoid haemorrhage, and subdural and epidural hematoma, but excluded isolated linear skull fractures.
Based on this set of predictors, the CHIP-prediction rule was previously developed for the identification of intracranial traumatic findings on CT, using logistic regression for the statistical modelling [1]. We compared the logistic regression model to alternative modelling techniques in developing prediction rules for intracranial findings on CT. We used the predictors listed in Table 2.

### Description of the modelling techniques

The alternative modelling techniques compared in this study are briefly described below [7].

*Bayes network*

A Bayesian network is a graphical model that displays variables (often referred to as nodes) in a dataset and the probabilistic, or conditional, dependencies between them. Causal relationships between nodes may be represented by a Bayesian network; however, the links in the network (also known as arcs) do not necessarily represent direct cause and effect. For example, a Bayesian network can be used to calculate the probability of a patient having a specific disease, given the presence or absence of certain symptoms and other relevant data, if the probabilistic dependencies between symptoms and disease as displayed on the graph hold true. Networks are robust to missing information and aim to make the best possible prediction using whatever information is present.
There are several reasons to use a Bayesian network:

- It helps to learn about (potentially causal) relationships.
- The network provides an efficient approach to prediction by parsimonious modelling and aims to avoid overfitting of data.
- It offers a clear visualization of the relationships involved.

*Neural net*

A neural network, sometimes called a multilayer perceptron, is a simplified model of the way the human brain processes information. It works by simulating a large number of interconnected simple processing units that resemble abstract versions of neurons. The processing units are arranged in layers. There are typically three parts in a neural network: an input layer, with units representing the predictor variables, one or more hidden layers and an output layer, with a unit representing the outcome variable.

The units are connected with varying connection strengths or weights. Input data are presented to the first layer, and values are propagated from each neuron to every neuron in the next layer. Eventually, a prediction is delivered from the output layer. The network learns by examining individual records, generating a prediction for each record and making adjustments to the weights whenever it makes an incorrect prediction. This process is repeated many times, and the network continues to improve its predictions until one or more of the stopping criteria have been met.

With the default setting, the network will stop training when the network appears to have reached its optimally trained state (90% accuracy). The networks that fail to train well are discarded as training progresses.

Initially, all weights are random, and the predictions that come out of the net are nonsensical. The network learns through training. Records for which the output is known are repeatedly presented to the network, and the predictions it gives are compared to the known outcomes.

As training progresses, the network becomes increasingly accurate in replicating the known outcomes. Once trained, the network can be applied to future patients for whom the outcome is unknown.

*CHAID*

The Chi-squared Automatic Interaction Detection model is a classification method for building decision trees by using chi-square analysis to identify optimal splits. CHAID first examines the cross tables between each of the predictor variables and the outcome and tests for significance using a chi-square test. If more than one of these relations is statistically significant, CHAID will select the predictor that has the smallest p-value. If a predictor has more than two categories, these are compared, and categories that show a similar outcome are collapsed together. This is done by successively joining the pair of categories showing the least significant difference. This category-merging process stops when all remaining categories differ at the specified testing level. For set predictors, any categories can be merged. For an ordinal set, only contiguous categories can be merged. Exhaustive CHAID is a modification of CHAID that more thoroughly examines all possible splits for each predictor but takes longer to compute. CHAID can generate non-binary trees, meaning that some splits have more than two branches. It

therefore tends to create a wider tree than the binary growing methods. CHAID works for all types of predictors.

*Support vector machine*

A Support Vector Machine (SVM) performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases from different classes. It claims to be a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. A SVM may particularly be suited to analyze data with large numbers of predictor variables. SVM has applications in many disciplines, including customer relationship management (CRM), image recognition, bioinformatics, text mining concept extraction, intrusion detection, protein structure prediction, and voice and speech recognition.

*CART*

The Classification And Regression Tree model is a tree-based classification and prediction model. The model uses recursive partitioning to split the training records into segments with similar output variable values. The modelling starts by examining the input variables to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two further subgroups and so on, until the stopping criterion is met.

*Decision list*

A Decision list model identifies subgroups or segments that show a higher or lower likelihood of a binary outcome relative to the overall sample. The model consists of a list of segments, each of which is defined by a rule that selects matching records. A given rule may have multiple conditions. Rules are applied in the order listed, with the first matching rule determining the outcome for a given record. Taken independently, rules or conditions may overlap, but the order of rules resolves ambiguity. If no rule matches, the record is assigned to the remainder rule.

**Cut-off values**

For each model we determined cut-off values and classification rules to achieve a sensitivity >0.95. To this end, we varied the cut-off values for each model from 0.015 to 0.05. Furthermore, the reduction in CT scans was calculated given a certain cut-off value. Reduction was defined as the percentage of subjects who would not undergo CT scanning since absence of intracranial findings on CT was predicted.

### Modelling

For the various modelling techniques we used Clementine Modeller version 12.0 in combination with SPSS 16.0. The comparison was made using performance characteristics including the area under the ROC curve, sensitivity and specificity. We used default modelling settings as far as possible (Appendix 1). For the CART model, however, we used an extended setting besides the default setting. The stopping criteria for the default setting were: 100 records in the parent branch and 50 records in the child branch. The stopping criteria for the extended setting were: 11 records in the parent branch and 10 records in the child branch. In both variants we used pruning (Appendix 2).

### Cross-validation

The models were validated using 10x10 cross-validation. The file was split into 10 random deciles. Each model was trained repeatedly on 9 deciles with predictions generated for the remaining decile. The AUC-values were calculated for the 10 training parts and the full set of 10 deciles which were left out of the training parts. The difference defined the optimism of each model, and this process was repeated 10 times. The optimism was subtracted from the apparent AUC-value for each model on the original sample to obtain optimism-corrected estimates of model performance [8].

## 2.3 RESULTS

### Comparison of the performance of the models

The logistic regression and CART models showed limited optimism in the AUC-values (<0.040, Table 3). The support vector machine model had a remarkably high optimism (0.171). The logistic regression model had the best performance (optimism-corrected AUC 0.787), followed by the Bayes network model (AUC 0.744) and the neural net model (AUC 0.726). The CHAID model and the decision list model had AUC-values of 0.699 and 0.634 respectively. The support vector machine model and the default CART model performed poorly with AUC-values 0.581 and 0.560 respectively. Although the CHAID model was more overfitted, the optimism-corrected AUC-value was much better than the CART analyses (Table 3).

The default CART model showed less statistical optimism than the extended CART model (0.008 versus 0.039 respectively). However, the optimism-corrected AUC-value was worse for the default CART model (AUC 0.560 versus 0.618 respectively, Table 3).

The logistic regression model had a sensitivity of 0.98 and a reduction of 20% at a cut off value of 0.02. The Bayes network model had a sensitivity of 0.97 and a reduction of 23% at a cut off value of 0.015. For the neural net model, it was not possible to achieve a sensitivity >0.95.

**Table 3** AUC-values

| Model | AUC | 95% CI for AUC | Mean AUC training | Mean AUC test | Optimism | Optimism-Corrected AUC |
|---|---|---|---|---|---|---|
| Logistic regression | 0.800 | 0.769-0.830 | 0.789 | 0.772 | 0.017 | 0.783 |
| Neural net | 0.782 | 0.751-0.814 | 0.785 | 0.746 | 0.038 | 0.744 |
| Bayes network | 0.806 | 0.777-0.836 | 0.808 | 0.743 | 0.065 | 0.741 |
| CHAID | 0.759 | 0.724-0.794 | 0.761 | 0.686 | 0.075 | 0.684 |
| Decision list | 0.674 | 0.633-0.715 | 0.673 | 0.626 | 0.048 | 0.627 |
| CART extended | 0.657 | 0.616-0.699 | 0.599 | 0.559 | 0.040 | 0.617 |
| Support vector machine | 0.754 | 0.714-0.794 | 0.740 | 0.578 | 0.162 | 0.592 |
| CART default | 0.568 | 0.527-0.609 | 0.556 | 0.537 | 0.019 | 0.549 |

## Graphical representations

The CART model is presented as a tree. The default CART model consisted of two predictor variables (Fracture skull and Cause), which were presented with three end nodes (Figure 1). The extended CART model consisted of six predictor variables (Fracture skull, EMV change, Cause, Memory deficit and Age per decade) presented in a tree with nine end nodes (Figure 2).

The Bayes network model is presented an interaction graph. It shows the relative importance of the predictors (Figure 3). The variable 'intracranial lesions' had a direct relation with the variable 'fracture skull' and the variable 'seizure'. It also showed a relation between the variable 'fracture skull' and the variable 'seizure'.

**Figure 1** CART model default

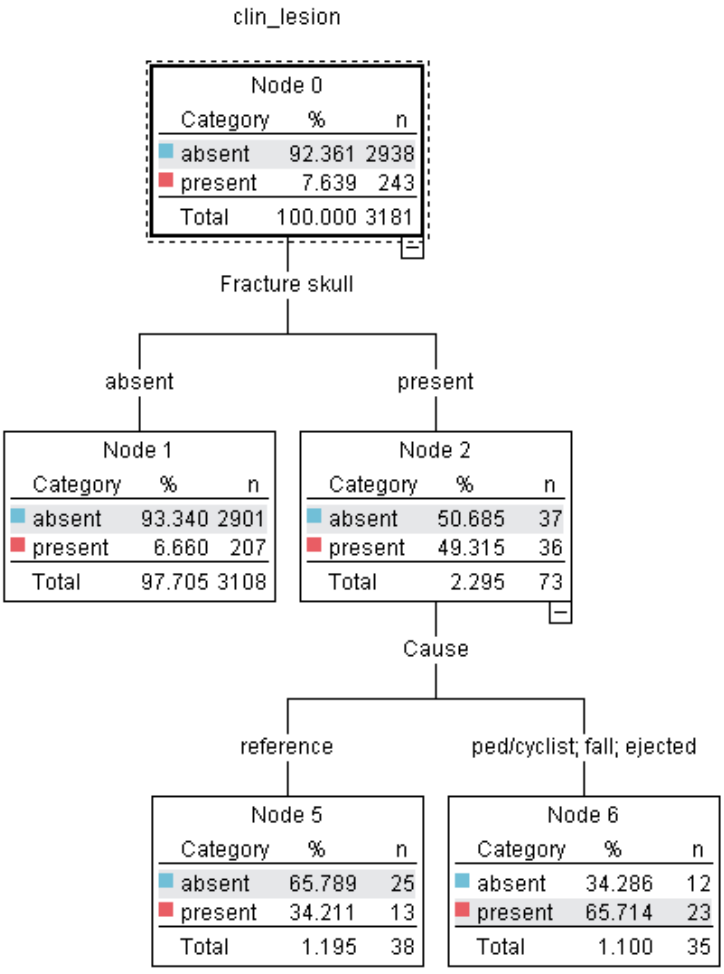


**Figure 2** CART model extended

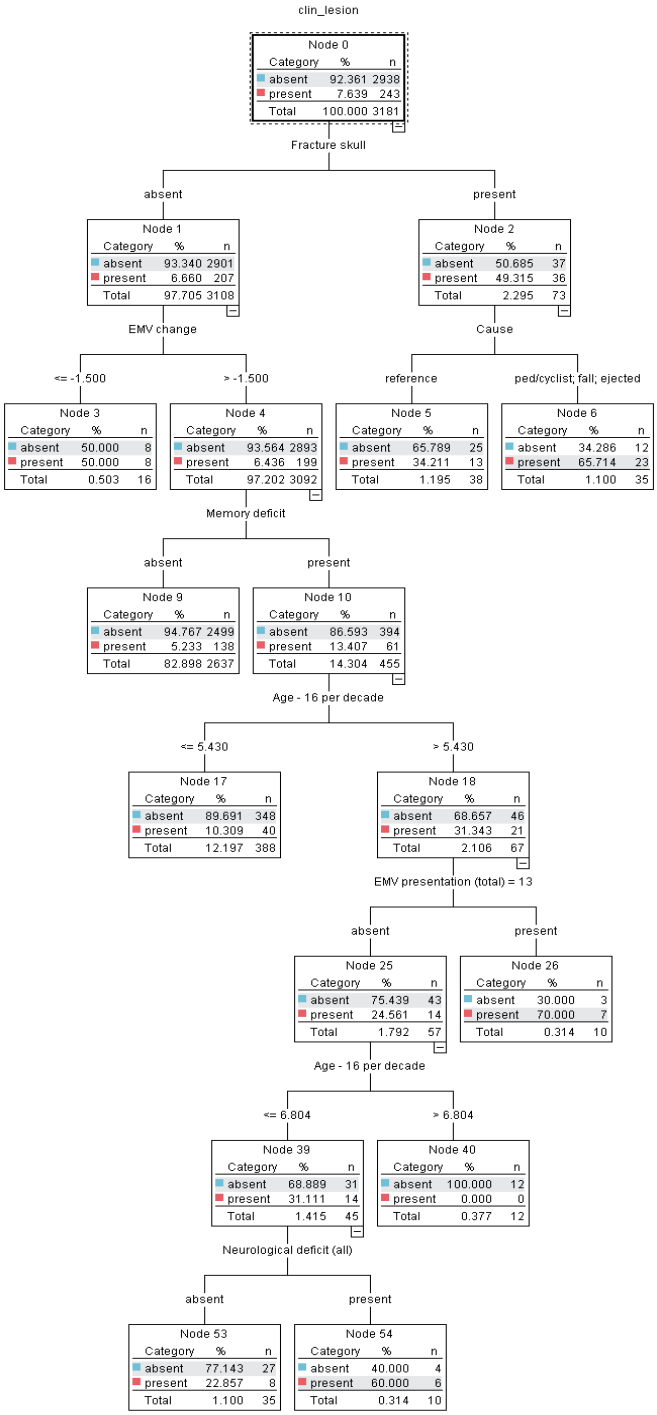


**Figure 3** Bayesian network model

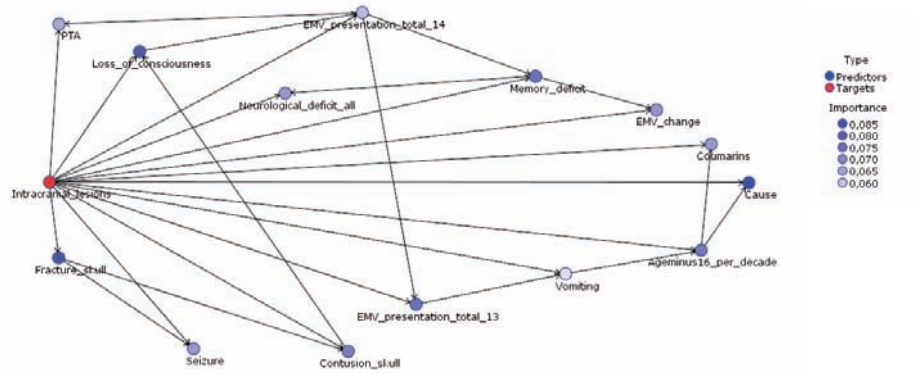



**Figure 4** Conditional probabilities of Intracranial lesions

| Probability | |
|---|---|
| 1 | 0 |
| 0.076 | 0.923 |

**Figure 5** Conditional probabilities of Fracture skull

| Parents | Probability | |
|---|---|---|
| Intracranial_lesions | 1 | 0 |
| 1 | 0.148 | 0.851 |
| 0 | 0.012 | 0.987 |

**Figure 6** Conditional probabilities of Seizure

| Parents | | Probability | |
|---|---|---|---|
| Fracture_skull | Intracranial_lesions | 1 | 0 |
| 1 | 1 | 0.083 | 0.916 |
| 1 | 0 | 0.027 | 0.972 |
| 0 | 1 | 0.009 | 0.990 |
| 0 | 0 | 0.005 | 0.994 |

The Bayes network model also presented the conditional probabilities (Figures 4, 5 and 6). Figure 6 shows that if fracture skull is absent and intracranial lesions are absent, the probability that seizure is absent equals 0.994.

Using Bayes theorem and the conditional probabilities in the figures 4, 5 and 6, we calculated that if seizure is absent, the predicted probability that intracranial traumatic findings are absent equals 92.5% (Figure 7).

**Figure 7** Calculation example

In general :

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(A \cap B \cap C) + P(A \cap \overline{B} \cap C)}{P(A \cap B \cap C) + P(\overline{A} \cap B \cap C) + P(A \cap \overline{B} \cap C) + P(\overline{A} \cap \overline{B} \cap C)} =$$

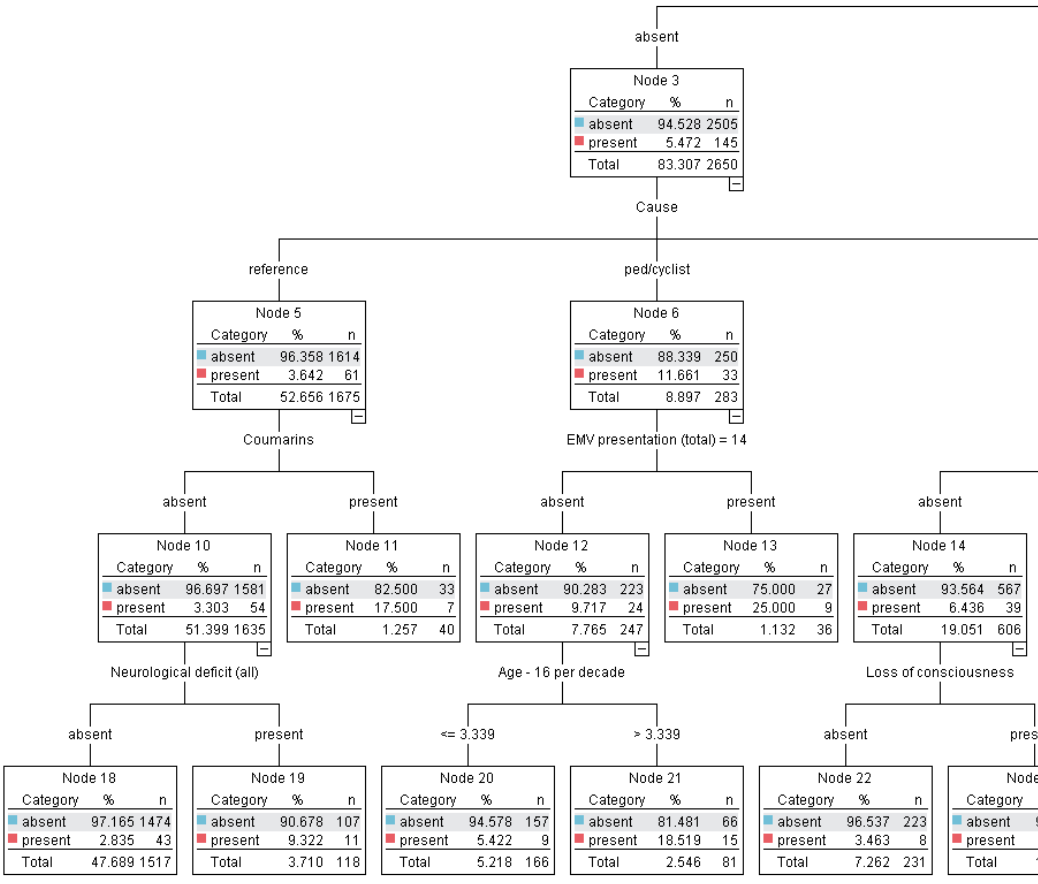
$$\frac{P(C|A \cap B) * P(B|A) * P(A) + P(C|\overline{A} \cap B) * P(B|\overline{A}) * P(\overline{A}) + P(C|A \cap \overline{B}) * P(\overline{B}|A) * P(A) + P(C|\overline{A} \cap \overline{B}) * P(\overline{B}|\overline{A}) * P(\overline{A})}{P(C|A \cap B) * P(B|A) * P(A) + P(C|\overline{A} \cap B) * P(B|\overline{A}) * P(\overline{A}) + P(C|A \cap \overline{B}) * P(\overline{B}|A) * P(A) + P(C|\overline{A} \cap \overline{B}) * P(\overline{B}|\overline{A}) * P(\overline{A})}$$

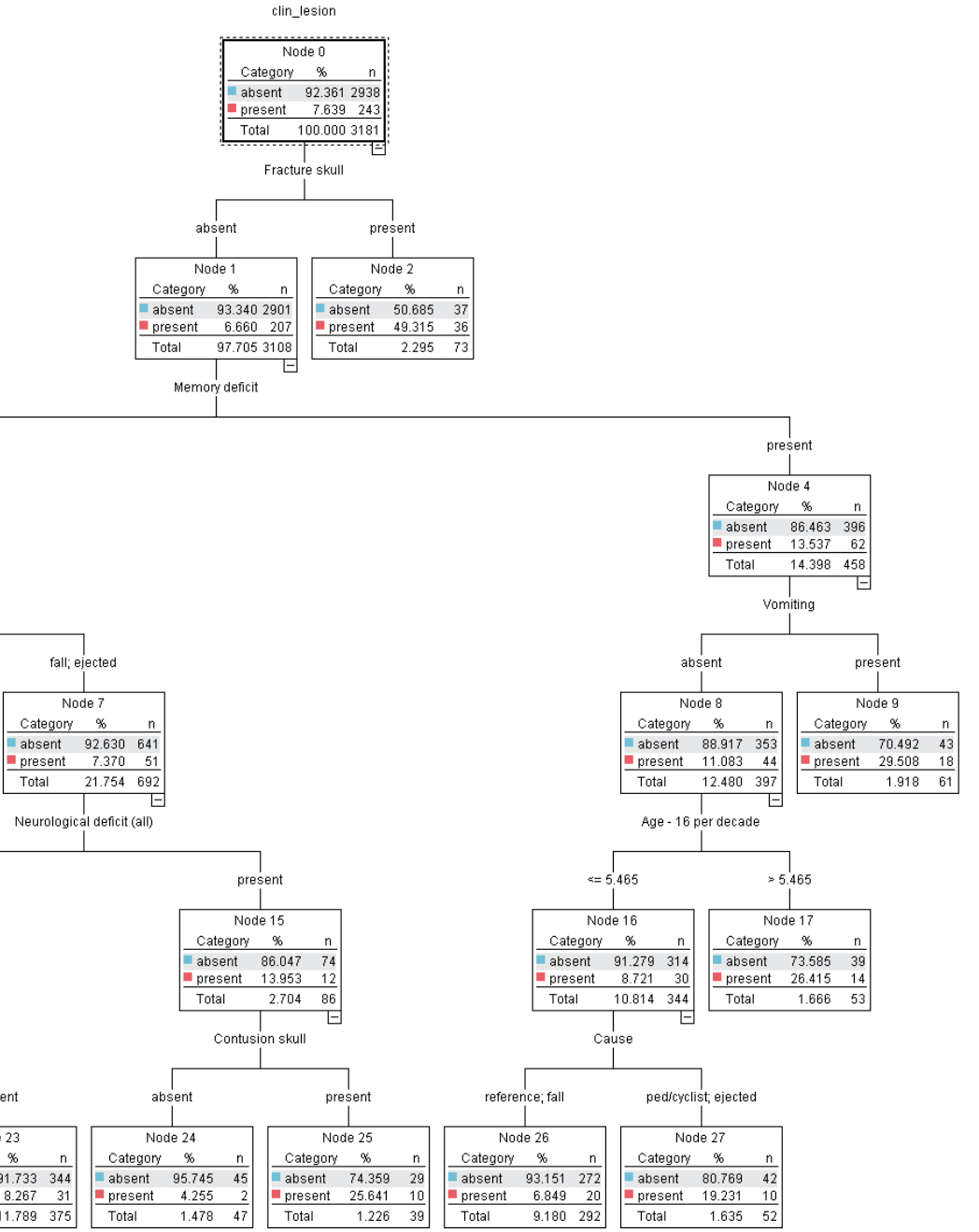
With A = Intracranial lesions (0 = absent, 1 = present), B = Fracture skull (0 = absent, 1 = present) and C = Seizure (0 = absent, 1 = present)

$$P(A = 0|C = 0) = \frac{0.994 * 0.987 * 0.923 + 0.972 * 0.012 * 0.923}{0.994 * 0.987 * 0.923 + 0.972 * 0.012 * 0.923 + 0.990 * 0.851 * 0.076 + 0.916 * 0.148 * 0.076} = 0.925$$

The CHAID model presented a tree graph. The tree consisted of fifteen end nodes and therefore of fifteen decision rules (Figure 8). Hence the tree size was much larger than that of the CART analyses (Figure 1 and Figure 2).

**Figure 8** CHAID model




clin_lesion
Node 0
Category % n
absent 92.361 2938
present 7.639 243
Total 100.000 3181
Fracture skull
absent
present
Node 1
Category % n
absent 93.340 2901
present 6.660 207
Total 97.705 3108
Node 2
Category % n
absent 50.685 37
present 49.315 36
Total 2.295 73
Memory deficit
present
Node 4
Category % n
absent 86.463 396
present 13.537 62
Total 14.398 458
Vomiting
absent
present
fall; ejected
Node 7
Category % n
absent 92.630 641
present 7.370 51
Total 21.754 692
Node 8
Category % n
absent 88.917 353
present 11.083 44
Total 12.480 397
Node 9
Category % n
absent 70.492 43
present 29.508 18
Total 1.918 61
Neurological deficit (all)
Age - 16 per decade
present
<= 5.465
> 5.465
Node 15
Category % n
absent 86.047 74
present 13.953 12
Total 2.704 86
Node 16
Category % n
absent 91.279 314
present 8.721 30
Total 10.814 344
Node 17
Category % n
absent 73.585 39
present 26.415 14
Total 1.666 53
Contusion skull
Cause
absent
present
reference; fall
ped/cyclist; ejected
Node 24
Category % n
absent 95.745 45
present 4.255 2
Total 1.478 47
Node 25
Category % n
absent 74.359 29
present 25.641 10
Total 1.226 39
Node 26
Category % n
absent 93.151 272
present 6.849 20
Total 9.180 292
Node 27
Category % n
absent 80.769 42
present 19.231 10
Total 1.635 52

## Presentation of the logistic regression model

The coefficients of the logistic regression model are presented in Table 4. The probabilities were calculated using Formula 1.

**Formula 1** Calculation probabilities logistic regression model (π)

$$\pi = \frac{1}{1+e^{-\Sigma x_i * b_i}}$$

**Table 4** Regression coefficients logistic model

| Variables | X | b |
|---|---|---|
| Fracture skull | Present | 2.34 |
| | Absent | 0.00 |
| EMV presentation (total) = 13 | Present | 1.37 |
| | Absent | 0.00 |
| EMV presentation (total) = 14 | Present | 0.72 |
| | Absent | 0.00 |
| Memory deficit | Present | 0.41 |
| | Absent | 0.00 |
| Contusion skull | Present | 0.59 |
| | Absent | 0.00 |
| Loss of consciousness | Present | 0.60 |
| | Absent | 0.00 |
| Seizure | Present | 0.84 |
| | Absent | 0.00 |
| Vomiting | Present | 0.88 |
| | Absent | 0.00 |
| Coumarins | Present | 0.87 |
| | Absent | 0.00 |
| Neurological deficit (all) | Present | 0.40 |
| | Absent | 0.00 |
| EMV change | EMV change | −0.32 |
| Cause | Reference | 0.00 |
| | pedastrian or cyclist | 1.27 |
| | Fall | 0.55 |
| | Ejected | 1.13 |
| Age - 16 per decade | Age - 16 per decade | 0.17 |
| PTA | <=2 hrs | 0.00 |
| | >2 hrs and <=4 hrs | 0.48 |
| | >4 hrs | 2.01 |
| Constant | Constant | −4.77 |

## 2.4 DISCUSSION

We found that alternative modelling techniques did not deliver better results in terms of applicability and performance in developing prediction rules for intracranial findings in patients with minor head injury than modelling based on conventional modelling techniques such as logistic regression. The performance of logistic regression was compared with six alternative modelling techniques using standard measures, specifically the receiver operating characteristic (ROC) curve. In a ROC curve, the trade-off between sensitivity and specificity is shown based on consecutive cut-off values. The key characteristic for model comparisons is the area under the ROC curve, which is equivalent to the concordance (or 'c') statistic.

The apparent AUC-values of each model were corrected for optimism using 10x10 cross-validation. Only the logistic regression model, the Bayes network model and the neural net model had satisfactory AUC-values (> 0.7), although it was impossible to achieve a sensitivity >0.95 for the neural net model. The CHAID model and the decision list model had AUC-values of 0.699 and 0.634 respectively, and the support vector machine model and the default CART model performed poorly (AUC-values <0.6).

At a cut-off value of 0.015, the logistic regression model would miss only 1% of the patients with intracranial traumatic findings (sensitivity 99%), whereas the Bayes network model would miss 3% (sensitivity 97%) at this cut-off. On the other hand, at this cut-off value the specificity of the Bayes model would be better (25%), and could potentially reduce the number of CT scans ordered by 23%. In contrast, the logistic regression model would only have 8% specificity and would reduce the number of CT scans ordered by 8% at a cut-off of 0.015. This illustrates the difficult trade-off between missing patients with intracranial traumatic findings versus the wish to reduce unnecessary CT scans in those without intracranial traumatic findings.

No modelling technique outperformed the relatively simple logistic regression model in terms of the optimism-corrected AUC-value. These findings may be seen as confirming the validity of the previously developed CHIP prediction rule [1]. However, it should be noted that these results are an internal validation of the developed CHIP-rule and that external validation is still required.

Our findings are in contrast to a recent study that advocated CART modelling to develop a prediction rule for CT scanning in children [4]. This can potentially be explained by the fact that modelling techniques such as CART are 'data hungry'. Therefore CART modelling may have been suitable for the Kuppermann study, which included 42,411 patients (376 with abnormal CT scans). However, it was not suitable for the CHIP database, which included only 3,181 patients (243 with abnormal CT scans). Also, the specific algorithm used in the Kuppermann study may have been different from the algorithm used in our study.

The superior performance of the logistic regression modelling might be explained by the high number of categorical variables (10 out of 14), which might favour logistic regression modelling. The somewhat disappointing performance of tree models like CHAID and CART may be more realistic, because these models are well suited for dealing with categorical and continuous variables, although the latter are categorized by these models.

Although the examined modelling techniques did not outperform logistic regression analysis, we can see a role for these techniques in providing a deeper insight into the interrelationships between predictors and outcome. For example, the Bayes network offered the advantage of showing a graphical representation of the direct relationships between the predictor variables and the outcome variable, as well as the first-order interactions. The CHAID model offered a tree graph which might give researchers insight into relevant risk groups. The neural net model, on the other hand, did have a satisfactory optimism-corrected AUC-value, but did not provide further insight into the medical problem. This alternative modelling technique has a black box character, which is a serious drawback for application in medical practice.

The outcomes of this study suggest that the use of alternative modelling techniques may also have practical value in ascertaining variables of critical import and in streamlining current existing guidelines. Smits et al. used 14 variables for their modelling based on expert opinion and previous studies. We started out with these same 14 variables to be able to compare the model of Smits et al. with modelling based on alternative modelling techniques. However, the CHAID model only used 10 out of these 14 variables. The variables PTA, Change, EMV-13 and Seizure were not used, which suggests that these variables may be of lower importance for the outcome. However, the CHAID model performed poorly in comparison with logistic regression modelling. For most of the evaluated models, the variables of critical import were: Fracture skull (v69), Cause (cause3) and Age - 16 per decade (age10). Based on our study, the guidelines should certainly contain these variables.

A priori, it is not fully predictable whether an alternative modelling technique will perform better than conventional modelling techniques. This depends on the internal structure of the prediction problem and on the characteristics of the modelling techniques. For example, tree modelling is well suited for a situation with many interactions between predictors, which might be missed with a default main effects logistic model. Neural nets are even more flexible in capturing interactions and non-linearities, which might be missed by other modelling techniques. It has been suggested that the balance between signal and noise is relatively unfavourable in many medical applications, making relatively simple regression models perform quite reasonably [9].

All these models can easily be evaluated, because capacity limitations for computer calculations no longer exist nowadays. The required software for evaluating the per-

formance of alternative modelling techniques is readily available (e.g. Clementine, R software, etc). The methods we used in this study may be applied to other studies using characteristics such as AUC-values, sensitivity and specificity. Internal validation can be performed using 10x10 cross-validation. From there, optimism-corrected AUC-values can readily be calculated.

Depending on the software used, it is possible to use the default setting or to choose an expert setting for the CART modelling. A researcher may use an expert setting for the number of levels below the root of a tree, for the number of records in the parent node and the child node, for applying or not applying pruning, for using weights for the categories of the outcome variable (costs) and so on. In our study, we used the default settings for the modelling as far as possible. Only in the evaluation of the CART model did we use an extended setting besides the default setting in order to achieve a higher AUC-value, but even then the performance of this model was poor.

In view of the applicability and simplicity of a prediction model, medical experts and researchers usually prefer a small number of predictors. However, this study shows that a considerable number of variables may be necessary to make an informed decision or a prediction with a high level of accuracy. The CHIP rule included 14 variables as major and minor risk factors, which all turned out to be indispensable.

By comparison, the default CART model appeared attractive, as it consisted of only 3 end nodes and therefore of 3 decision rules. Unfortunately, this model showed a poor performance.

Larger models may lead to better performance when all predictors are in fact predictive of the outcome [10]. While the number of predictors should therefore not be unduly limited, the applicability and simplicity of a decision rule might still be improved by using a model that provides a clarifying presentation of all the relevant variables and their mutual dependencies. Therefore the search for superior models with attractive presentation formats should continue.

## Conclusions

No alternative modelling technique outperformed the logistic regression model. However, the Bayes network model had a presentation format which provided more detailed insights into the structure of the prediction problem. The search for methods with good predictive performance and an attractive presentation format should continue.

## 2.5 APPENDIX 1 MODELLING SETTINGS

| Bayes network | |
|---|---|
| Build Settings | |
| | Use partitioned data: false |
| | Variable importance.LABEL: true |
| | Calculate raw propensity scores: true |
| | Calculate adjusted propensity scores: false |
| | Use frequency field: false |
| | Continue training existing model: false |
| | Structure type: TAN |
| | Include feature selection preprocessing step: false |
| | Parameter learning method: Maximum likelihood |
| | Mode: Simple |
| | Use only complete records: true |
| | Append all probabilities: false |
| | Independence test: Likelihood ratio |
| | Significance level: 0,01 |
| | Maximal conditioning set size: 5 |
| | Inputs always selected: [] |
| | Maximum number of inputs: 10 |

| Neural net | |
|---|---|
| Build Settings | |
| | Use partitioned data: false |
| | Calculate variable importance: true |
| | Calculate raw propensity scores: true |
| | Calculate adjusted propensity scores: false |
| | Method: Quick |
| | Stop on: Default |
| | Set random seed: true |
| | Set random seed: true |
| | Prevent overtraining: false |
| | Sample %: 50,0 |
| | Optimize: Memory |
| | Mode: Simple |
| Analysis | |
| | Estimated accuracy: 93,587 |
| | Input Layer: 29 neurons |

| | |
|---|---|
| | Hidden Layer 1: 3 neurons |
| | Output Layer: 1 neurons |

| CHAID | |
|---|---|
| Analysis | |
| | Tree depth: 5 |
| Build Settings | |
| | Use partitioned data: false |
| | Calculate variable importance: true |
| | Calculate raw propensity scores: true |
| | Calculate adjusted propensity scores: false |
| | Use frequency: false |
| | Use weight: false |
| | Levels below root: 5 |
| | Mode: Simple |
| | Use misclassification costs: false |

| Support vector machine | |
|---|---|
| Build Settings | |
| | Use partitioned data: false |
| | Variable importance.LABEL: true |
| | Calculate raw propensity scores: true |
| | Calculate adjusted propensity scores: false |
| | Mode: Simple |
| | Append all probabilities (valid only for categorical targets): false |
| | Stopping criteria: 1.0E-3 |
| | Kernel type: RBF |
| | Regularization parameter (C): 10 |
| | Regression precision (epsilon): 0,1 |
| | RBF gamma: 0,1 |
| | Gamma: 1,0 |
| | Bias: 0,0 |
| | Degree: 3 |

| CART default | |
|---|---|
| Analysis | |
| | Tree depth: 2 |
| Build Settings | |

| | |
|---|---|
| | Use partitioned data: false |
| | Calculate variable importance: true |
| | Calculate raw propensity scores: true |
| | Calculate adjusted propensity scores: false |
| | Use frequency: false |
| | Use weight: false |
| | Levels below root: 10 |
| | Mode: Simple |
| | Use misclassification costs: false |

| CART extended | |
|---|---|
| Build Settings | |
| | Use partitioned data: false |
| | Calculate variable importance: true |
| | Calculate raw propensity scores: true |
| | Calculate adjusted propensity scores: false |
| | Use frequency: false |
| | Use weight: false |
| | Levels below root: 5 |
| | Mode: Expert |
| | Maximum surrogates: 5 |
| | Minimum change in impurity: 0,0 |
| | Impurity measure for categorical targets: Gini |
| | Stopping criteria: Use absolute value |
| | Minimum records in parent branch: 11 |
| | Minimum records in child branch: 10 |
| | Prune tree: true |
| | Use standard error rule: false |
| | Prior probabilities: Based on training data |
| | Adjust priors using misclassification costs: false |
| | Use misclassification costs: false |

| Decision list | |
|---|---|
| Build Settings | |
| | Use partitioned data: false |
| | Calculate raw propensity scores: true |
| | Calculate adjusted propensity scores: false |
| | Use frequency: false |
| | Target value: 1,0 |

| | |
|---|---|
| | Search direction: Up |
| | Maximum number of segments: 5 |
| | Minimum segment size (as percentage): 5,0 |
| | Minimum segment size (as absolute value): 50 |
| | Maximum number of attributes: 5 |
| | Allow attribute re-use: true |
| | Confidence interval for new conditions (%): 95,0 |
| | Mode: Simple |

| Logistic regression | |
|---|---|
| Build Settings | |
| | Use partitioned data: false |
| | Calculate variable importance: true |
| | Calculate raw propensity scores: true |
| | Procedure: Multinomial |
| | Base category: 0 |
| | Model type: Main Effects |
| | Include constant in equation: true |
| | Mode: Simple |
| | Multinomial Method: Enter |

## 2.6 APPENDIX 2 CHARACTERISTICS OF THE MODELS

| | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | Bayes network | CHAID and CART | Decision list | Support vector machine | Neural net | Logistic regression |
| Categorizing of continuous predictor variables | Yes | x | x | x | | x | |
| | No | | | | x | | x |
| Outcome | Continuous | | x | | x | x | |
| | Categorial | x | x | | x | x | |
| | Dichotomous | x | x | x | x | x | x |
| Iteractions | Assumed | | x | x | | x | |
| | Flexible | x | | | x | | |
| | Possible | | | | | | x |
| Selection of predictor variables | Assumed | x | x | x | | | |
| | Flexible | | | | x | x | x |
| Graphical output | Tree graph | | x | | | | |
| | Interaction graph | x | | | | | |
| | Variable importance | x | x | | x | x | |
| Formula | Yes | | | | | | x |
| | No | x | x | x | x | x | |

## REFERENCES


1. Smits M, Dippel DWJ, Steyerberg EW, de Haan GG, Dekker HM, Vos PE, Kool DR, Nederkoorn PJ, Hofman PAM, Twijnstra A, others: Predicting intracranial traumatic findings on computed tomography in patients with minor head injury: the CHIP prediction rule. *Ann Intern Med* 2007, 146:397–405.
2. Stiell IG, Wells GA, Vandemheen K, Clement C, Lesiuk H, Laupacis A, McKnight RD, Verbeek R, Brison R, Cass D, others: The Canadian CT Head Rule for patients with minor head injury. *Lancet* 2001, 357:1391–1396.
3. Stiell IG, Clement CM, Rowe BH, Schull MJ, Brison R, Cass D, Eisenhauer MA, McKnight RD, Bandiera G, Holroyd B, others: Comparison of the Canadian CT Head Rule and the New Orleans Criteria in patients with minor head injury. *Jama* 2005, 294:1511–1518.
4. Kuppermann N, Holmes JF, Dayan PS, Hoyle JD, Atabaki SM, Holubkov R, Nadel FM, Monroe D, Stanley RM, Borgialli DA, others: Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. *Lancet* 2009, 374: 1160–1170.
5. Breiman L, Friedman JH, Olshen RA, Stone CJ: *Classification and Regression Trees. Volume 19*; 1984.
6. Breiman L, Stone CJ: Parsimonious binary classification trees. *Technol Serv Corp St Monica, Calif Tech Rep TSCCSD-TN-004* 1978.
7. SPSS C: 8.0 User's Guide. *Integr Solut Ltd* 2003.
8. Steyerberg EW: *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Volume 19*; 2009.
9. Zani S: *Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6-8, 2005*. Springer; 2006.
10. Harrell FE: *Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer; 2001.

# Chapter 3

## Risk prediction with machine learning and regression methods

Ewout W. Steyerberg, Tjeerd van der Ploeg, Ben van Calster

Biom J. 2014 Jul;56(4):601-6. doi:10.1002/bimj.201300297.

Machine learning techniques may have important roles in the medical statistical literature. We aimed to reflect the pros and cons of machine learning techniques in comparison with traditional regression modelling for risk prediction. First, we note that considering non-linearity is essential in a modern approach to regression analysis. Second, we emphasize the use of penalization procedures for a fair comparison of regression to MLT. Next, we discuss the role of model uncertainty, which argues for relatively simple models, and the balance between information from outside the data under study versus what can be learned from the data. We end with a discussion on the potential role of MLT in addition to regression modelling.

Key words: Machine learning; Prediction; Regression

## 3.1 INTRODUCTION

Cross-fertilization between medical statistics and epidemiology on the one hand and machine learning techniques (MLT) on the other can be very stimulating (Kruppa et al., 2014a; Kruppa et al., 2014b). Not only is probability estimation discussed for dichotomous outcomes, but also for multi-category (or polytomous, multinomial) outcomes, which is an underdeveloped research area (Van Calster et al., 2012b). Probability estimation is key to the area of risk prediction, which is growing in importance in medicine, where personalized medicine becomes more and more possible through the combination of classical risk predictors and biomarkers.

The first paper focuses on theoretical aspects, such as consistency of probability estimation (Kruppa et al., 2014a). For example, for the Nearest Neighbor (NN) method the authors report that the error in the estimation of probabilities converges to zero if the sample size tends to infinity, while this is not strictly true for Random Forests (RF). Consistency does not hold for logistic regression (logreg), where the validity of probability estimates depends on the model specification. Simulation studies are provided which show that each of these methods can fail to provide reasonable predictions. Calibration properties were particularly poor for some variants of Support Vector Machines (SVMs) in some simulations. Various performance criteria were studied, specifically squared scoring rules such as the Brier score. Rank-based measures such as the area under the ROC curve were also used, for which extensions to multi-category evaluation have recently been proposed, such as the Polytomous Discrimination Index (Van Calster et al., 2012a). Likelihood based performance measures might also have been be used, such as Nagelkerke's R2 (Austin & Steyerberg, 2013), but these would probably have led to the same impression of performance. Finally, the paper nicely illustrates that some methods behave very similarly, e.g. two variants of NN, and SVM with linear kernel and logreg.

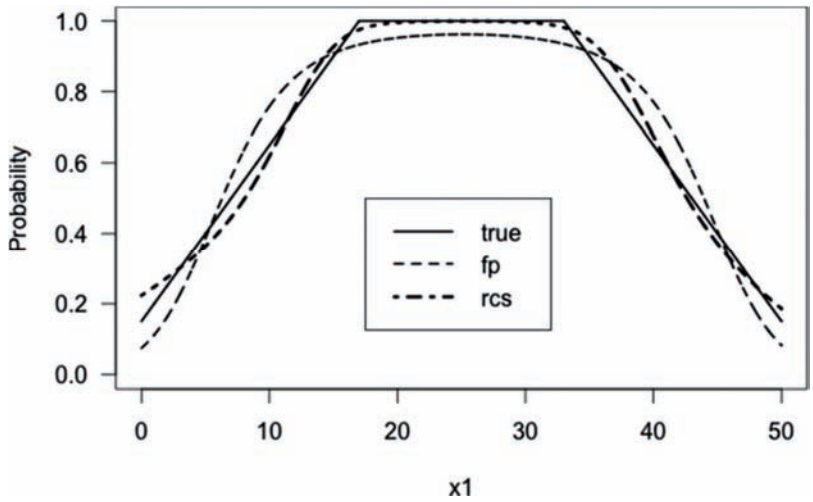
Below we first discuss tuning and implementation aspects of MLT and regression models (section 2), followed by reflections on model uncertainty (section 3) and a possible sensible modelling strategy (section 4). We discuss the potential role of MLT in addition to regression modelling in section 5.

## 3.2 TUNING, TRADITIONS, AND MODERN APPROACHES IN REGRESSION MODELLING

As emphasized by the authors, one issue of attention with machine learning techniques (MLT) is that they have various tuning parameters, such as the number of neighbors to consider in NN, the regularization parameters and type of kernel for SVM, and the tree

specifications for RF, which essentially serve to control the complexity of the fitted model. Similarly, various strategies and modelling approaches are possible for logreg. First, prediction modelers of medical data should assess non-linearity of continuous variables (Steyerberg, 2009). The blind application of the logistic regression model y~x1+x2, as was presented in Simulation I, is divorced from reality. The underlying circle model requires some kind of increasing and decreasing functions for x1 and x2. Any epidemiologist would do some form of data inspection, and would immediately note the more or less squared relation with x (Figure 1). After that no one would model linear effects for x1 and x2. Preferences for modelling non-linearity vary: Harrell advisesrestricted cubic splines (rcs) as a default tool in regression modelling (Harrell, 2001), while Royston & Sauerbrei (2008) advocate the use of fractional polynomials (FP). For illustration, we fitted FP and rcs functions in a simulation with 5000 subjects (Figure 1, using R packages mfp and rms). The true effect of x1 is a linear increase from x1=0 to x1=17, a probability of 1 between x1=17 and x1=33, and a linear decrease between x1=33 and x1=50. For the FP model, a linear term plus square term are selected for x1. This FP model follows the true shape well, although the probability of 1 is not reached, and low probabilities are underestimated. The rcs model (with 5 knots, 4 df) reached the plateau probability of 1, but slightly overestimated low probabilities at x1=0 and x1=50. The models y~fp(x1)+fp(x2) and y~rcs(x1)+rcs(x2) had Brier scores below 0.15, which is equivalent to the best performing MLT in this simulation (NN, SVM-Bessel).

**Figure 1:** The probability of y=1 in Simulation study I, for x2=25. We performed a simulation study of 5000 subjects, where the selected FP function was (x1+0.1)+(x1+0.1)^2. The rcs function used 5 knots (4 df).



So, as may be expected, a reasonably specified logreg model performs very well in simulation I.
Second, whereas some form of regularization is indispensable for MLT due to their flexibility, similar techniques exist for logreg to penalize or shrink model coefficients. Examples are simple shrinkage, L1 (LASSO) or L2 (ridge) penalization, or Bayesian approaches (Steyerberg, 2009). The LASSO method uses a L1 penalty to shrink regression coefficients to zero (Tibshirani, 1996). Hence LASSO combines variable selection with shrinkage while still providing adequate predictions, as observed in a large simulation study for patients with an acute myocardial infarction (Steyerberg et al., 2000). Similar to the improvement of RF over CART for prediction (Austin et al., 2012), we should use penalized rather than traditional approaches for logreg if comparisons are made between logreg and MLT.

## 3.3 MODEL UNCERTAINTY

The main problem for prediction models is model uncertainty. We can usually specify various models, which all reasonably describe the data (Breiman, 2001). In medical research, we may often have a relatively long list of potential predictors, e.g. 49 for Application I (stroke) (Kruppa et al., 2014b). This list was apparently based on solid grounds (a systematic literature review), but some reduction might have been possible by posing stricter criteria on the evidence underpinning a potential predictive effect, such as consistency of a substantial effect size across multiple studies. It is not plausible that a medical problem has 49 equally important predictors. For example, we identified only 3 key predictors of 6-month outcome in a systematic literature review for patients with traumatic brain injury (Mushkudiani et al., 2008). In this prediction problem, Age, Glasgow Coma Scale -especially the Motor component-, and pupillary reactivity strongly predicted 6-month mortality (Perel et al., 2008) (Steyerberg et al., 2008). Models with these key predictors performed well in temporal and geographical validations (Roozenbeek et al., 2012). Only minor improvements were noted by including other characteristics, such as CT scan findings, while many clinicians would consider these characteristics vital for prediction.
Moreover, it is well known that medical data typically have a poor signal to noise ratio for predictors. This has two implications. First, sample size and penalization are key factors to accurate prediction modelling. This is true for regression models, and even more so for MLT. MLT are more flexible than regression, which makes them more data hungry. A technique such as NN may be extreme in data requirements, because of its fully non-parametric nature. Second, simpler model specifications may often be sufficient to capture the main structure of a prediction problem. Extreme non-linearity such as

in the presented Simulation I is implausible in medical research. Complex higher order interactions may occasionally exist but are impossible to identify in reasonably sized medical data sets. This is supported by recent studies that report similar performance of logreg vs MLT (Van Calster et al., 2009) (Van Calster et al., 2010) (Van der Ploeg et al, 2011) (Austin et al., 2013).

## 3.4 SENSIBLE PREDICTION MODELLING IN MEDICAL DATA

Medical data sets are often of too small a size to be able to reliably address difficult research questions, such as determining which predictors are important and which are not. For example, reliably determining which of 49 characteristics predict mortality may require far larger numbers of events than occurring among the training set of 1737 patients in Application I (Kruppa et al., 2014b). In addition, backward elimination is a standard approach for variable selection in regression analysis, commonly requiring $p<0.05$ for predictors in a prediction model. Many drawbacks have been discussed in the past, including biased estimation of regression coefficients, distortion of the estimation of variance and p-values, and instability of the selected set of predictors (Austin & Tu, 2004) (Sauerbrei & Schumacher, 1992) (Steyerberg et al., 1999). For probability estimation the most relevant issue is that stepwise selection leads to suboptimal prediction: only the most prominent predictors are selected, so information from close-to-significant predictors is lost, and effects are exaggerated, which leads to too extreme predictions (Steyerberg et al., 2001).
Sensible modelling should find a balance between external knowledge from outside the data versus what can be learned from the data. The smaller the data set available, the more we have to rely on external information. This holds primarily for the list of candidate predictors in a model, which is relevant to both MLT and logreg. But it also holds for issues such as whether we should rely on the additivity assumption in logreg, i.e. whether we should consider statistical interaction terms. Some traditional statisticians might consider assessment of interactions as good modelling practice, while others would warn of overfitting by the potential for inclusion of spurious interactions. Findings in prior studies and sample size of the data under study are key considerations for such strategies (Steyerberg, 2009).

## 3.5 A ROLE FOR MLT IN ADDITION TO REGRESSION?

MLT have various attractive properties, such as their focus on regularization and on finding algorithms and classification models that work, rather than focusing strongly on

theory of an assumed stochastic data model (Breiman, 2001). Clinical risk prediction research uses a similar philosophy, focusing on performance issues such as discrimination, calibration, utility, and impact. Nevertheless MLT also have various problems. If we aim for an important role of prediction models in medicine, we need to follow a framework that not only includes model development, but involves a process of validation and updating of models (Steyerberg et al., 2013). Updating may require adjustments to local settings (van Houwelingen, 2000) (Steyerberg et al., 2004). In logreg, simple updating to the average probability is easily achieved by changing the model intercept, while this is difficult for MLT.

Furthermore, interpretability to a clinical audience is essential, as Kruppa et al. rightly notice. Logistic regression models can transparently be presented, with insight into the relative effects of predictors by odds ratios and in nomograms, score charts and other displays. Such presentations are not possible for MLT, although efforts to this end have been undertaken (Van Belle et al., 2012). On the other hand, we notice that prediction models are increasingly implemented on the internet. For example, a risk calculator for the probability of Lynch syndrome related mutation is accessed over 1000 times a month (Kastrinos et al., 2011). Web-based calculation of risk may allow the underlying model to be quite complex, e.g. an MLT.

Some characteristics of MLT and regression modelling techniques are summarized in Table 1. An NN approach may be attractive because of the theoretical property of consistency, but is data hungry (requires huge sample sizes) and lacks interpretability, similar to RF and SVM. The consistency of RF and SVM is not fully proven, but the flexibility is large. Although logreg is not consistent in the estimation of probabilities, the flexibility can be substantial with a modern modelling strategy. Naive fitting of linear main effects and automatic selection methods such as backward stepwise selection with $p<0.05$ are suboptimal implementations of logreg. Non-linear transformations can readily be made by rcs and FP functions, and the shrinkage or penalization methods such as LASSO provide better than standard predictive performance. Sample size requirements for logreg depend on how much external evidence is available, and how

**Table 1** Characteristics of MLT and regression modelling techniques.

| Method | Consistency | Flexibility | Sample size | Interpretability |
|---|---|---|---|---|
| NN | + | + | - | - |
| RF | +/- | + | +/- | - |
| SVM | +/- | + | +/- | - |
| Logreg | - | +/- | + | + |

NN: nearest neighbors; RF: random forest using probability estimation trees;
SVM: support vector machine; logreg: logistic regression

much the analyst is willing to rely on such evidence, e.g. on the relevance and effects of predictors. Interpretability of effect sizes is readily possible for a medically trained audience, and model updating can readily be achieved with simple or more advanced procedures.

All in all, we envision that logreg will remain the main modelling approach to probability estimation in medical risk prediction, especially when applied with modern approaches. MLT may have a supplementary role, in highly complex problems and to provide a comparison to regression results.

## REFERENCES


1. Austin, P.C. & Tu, J.V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol, 57, 1138-46.
2. Austin, P.C., Lee, D.S., Steyerberg, E.W. & Tu, J.V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? Biom J, 54, 657-73.
3. Austin, P.C. & Steyerberg, E.W. (2013). Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. Stat Med, 32, 661-72.
4. Austin, P.C., Tu, J.V., Ho, J.E., Levy, D. & Lee, D.S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol, 66, 398-407.
5. Breiman, L. (2001). Statistical modelling: the two cultures. Statist Sci, 16, 199-215.
6. Harrell, F.E. (2001). Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. Springer: New York.
7. Kastrinos, F., Steyerberg, E.W., Mercado, R., Balmana, J., Holter, S., Gallinger, S., Siegmund, K.D., Church, J.M., Jenkins, M.A., Lindor, N.M., Thibodeau, S.N., Burbidge, L.A., Wenstrup, R.J. & Syngal, S. (2011). The PREMM(1,2,6) model predicts risk of MLH1, MSH2, and MSH6 germline mutations based on cancer history. Gastroenterology, 140, 73-81.
8. Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I.R., Malley, J.D. & Ziegler, A. (2014a). Probability estimation with machine learning methods for dichotomous and multi-category outcome: Theory. Biom J.
9. Kruppa, J., Liu, Y., Diener, H.C., Holste, T., König, I.R., Weimar, C. & Ziegler, A. (2014b). Probability estimation with machine learning methods for dichotomous and multi-category outcome: Applications. Biom J.
10. Mushkudiani, N.A., Hukkelhoven, C.W., Hernandez, A.V., Murray, G.D., Choi, S.C., Maas, A.I. & Steyerberg, E.W. (2008). A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol, 61, 331-43.
11. Perel, P., Arango, M., Clayton, T., Edwards, P., Komolafe, E., Poccock, S., Roberts, I., Shakur, H., Steyerberg, E. & Yutthakasemsunt, S. (2008). Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. BMJ, 336, 425-9.
12. Roozenbeek, B., Lingsma, H.F., Lecky, F.E., Lu, J., Weir, J., Butcher, I., McHugh, G.S., Murray, G.D., Perel, P., Maas, A.I. & Steyerberg, E.W. (2012). Prediction of outcome after moderate and severe traumatic brain injury: external validation of the IMPACT and CRASH prognostic models. Crit Care Med, 40, 1609-17.

13. Royston, P. & Sauerbrei, W. (2008). Multivariable model-building : a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. John Wiley: Chichester, England.

14. Sauerbrei, W. & Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model. Stat Med, 11, 2093-109.

15. Steyerberg, E.W., Eijkemans, M.J. & Habbema, J.D. (1999). Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol, 52, 935-42.

16. Steyerberg, E.W., Eijkemans, M.J., Harrell, F.E., Jr. & Habbema, J.D. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med, 19, 1059-79.

17. Steyerberg, E.W., Eijkemans, M.J., Harrell, F.E., Jr. & Habbema, J.D. (2001). Prognostic modelling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making, 21, 45-56.

18. Steyerberg, E.W., Borsboom, G.J., van Houwelingen, H.C., Eijkemans, M.J. & Habbema, J.D. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med, 23, 2567-86.

19. Steyerberg, E.W., Mushkudiani, N., Perel, P., Butcher, I., Lu, J., McHugh, G.S., Murray, G.D., Marmarou, A., Roberts, I., Habbema, J.D. & Maas, A.I. (2008). Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med, 5, e165

20. Steyerberg, E.W. (2009). Clinical prediction models: a practical approach to development, validation, and updating. Springer: New York.

21. Steyerberg, E.W., Moons, K.G., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., Riley, R.D., Hemingway, H., Altman, D.G. & Group, P. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med, 10, e1001381.

22. Tibshirani, R. (1996). Regression and shrinkage via the Lasso. J R Stat Soc, Ser B, 58, 267-288.

23. Van Belle, V.M., Van Calster, B., Timmerman, D., Bourne, T., Bottomley, C., Valentin, L., Neven, P., Van Huffel, S., Suykens, J.A. & Boyd, S. (2012). A mathematical model for interpretable clinical decision support with applications in gynecology. PLoS One, 7, e34312.

24. Van Calster, B., Valentin, L., Van Holsbeke, C., Testa, A.C., Bourne, T., Van Huffel, S. & Timmerman, D. (2010). Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. BMC Med Res Methodol, 10, 96.

25. Van Calster, B., Van Belle, V., Vergouwe, Y., Timmerman, D., Van Huffel, S. & Steyerberg, E.W. (2012a). Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. Stat Med, 31, 2610-26.

26. Van Calster, B., Vergouwe, Y., Looman, C.W., Van Belle, V., Timmerman, D. & Steyerberg, E.W. (2012b). Assessing the discriminative ability of risk models for more than two outcome categories. Eur J Epidemiol, 27, 761-70.

27. Van Calster, B., Condous, G., Kirk, E., Bourne, T., Timmerman, D. & Van Huffel, S. (2009). An application of methods for the probabilistic three-class classification of pregnancies of unknown location. Artif Intell Med, 46, 139-54.

28. Van der Ploeg, T., Smits, M., Dippel, D.W., Hunink, M. & Steyerberg, E.W. (2011). Prediction of intracranial findings on CT-scans by alternative modelling techniques. BMC Medical Research Methodology, 11, 143.

29. Van Houwelingen, H.C. (2000). Validation. calibration, revision and combination of prognostic survival models. Stat Med, 19, 3401-15.

# Chapter 4

## Prediction of survival with alternative modelling techniques using pseudo values

Tjeerd van der Ploeg, Frank Datema, Robert Baatenburg de Jong, Ewout W. Steyerberg

PLoS One. 2014 Jun 20;9(6):e100234. doi: 10.1371/journal.pone.0100234.

## ABSTRACT

### Background

The use of alternative modelling techniques for predicting patient survival is complicated by the fact that some alternative techniques cannot readily deal with censoring, which is essential for analyzing survival data. In the current study, we aimed to demonstrate that pseudo values enable statistically appropriate analyses of survival outcomes when used in seven alternative modelling techniques.

### Methods

In this case study, we analyzed survival of 1282 Dutch patients with newly diagnosed Head and Neck Squamous Cell Carcinoma (HNSCC) with conventional Kaplan-Meier and Cox regression analysis. We subsequently calculated pseudo values to reflect the individual survival patterns. We used these pseudo values to compare recursive partitioning (RPART), neural nets (NNET), logistic regression (LR) general linear models (GLM) and three variants of support vector machines (SVM) with respect to dichotomous 60-month survival, and continuous pseudo values at 60 months or estimated survival time. We used the area under the ROC curve (AUC) and the root of the mean squared error (RMSE) to compare the performance of these models using bootstrap validation.

### Results

Of a total of 1282 patients, 986 patients died during a median follow-up of 66 months (60-month survival: 52% [95% CI:50%-55%]). The LR model had the highest optimism corrected AUC (0.791) to predict 60-month survival, followed by the SVM model with a linear kernel (AUC 0.787). The GLM model had the smallest optimism corrected RMSE when continuous pseudo values were considered for 60-month survival or the estimated survival time followed by SVM models with a linear kernel. The estimated importance of predictors varied substantially by the specific aspect of survival studied and modelling technique used.

### Conclusions

The use of pseudo values makes it readily possible to apply alternative modelling techniques to survival problems, to compare their performance and to search further for promising alternative modelling techniques to analyze survival time.

## 4.1 BACKGROUND

Predicting the survival probability of patients is important for various purposes in biomedical research, such as patient counselling, medical decision making, and benchmarking. The conventional analysis of survival problems mainly relies on Kaplan-Meier analysis and Cox regression modelling to predict the survival probability in relation to predictor variables [1] [2].

Alternative modelling techniques are available, such as support vector machines and artificial neural networks [3] [4] [5], which might possibly provide better predictions. For example, feed forward *neural networks* were already used in 1998 for the analysis of censored survival data [6]. In 2007, applications of random survival forests were described [7]. In 2009, prognostic indexes were compared using data mining techniques and Cox regression analysis in breast cancer data [8].

In 2000, Schwarzer and Vach [9] reviewed the use of artificial neural networks in medical research and found several problems. A major problem was that some of the alternative techniques did not deal adequately with censoring, which is essential for analyzing survival data. The conventional analysis of survival outcomes requires two variables: the status of the patient (e.g. dead or alive) and the time point at which this status is measured. In 2008, Klein et al. [10] [11] proposed to predict the survival at particular time points using pseudo values, which combine the variables status and time point in one outcome variable. The use of these pseudo values in generalized estimating equation modelling (GEE) using a log-minus-log link function leads to statistically appropriate analyses, which are in line with the results of Cox regression modelling.

In the current study, we aimed to study the use of pseudo values for analyses of survival outcomes with other modelling techniques, including support vector machines (SVM), neural networks (NNET), general linear models (GLM), recursive partitioning (RPART) and logistic regression (LR). To compare the performance, we applied these techniques and conventional regression analysis in the prediction of survival of 1282 Dutch patients with Head and Neck Squamous Cell Carcinoma (HNSCC), using predictors as described in earlier studies [12] [13] [14]. The survival of this particular population of newly diagnosed patients with HNSCC has already been studied by applying conventional Kaplan-Meier analysis, Cox regression and random survival forests (RSF) to 60-month survival and overall survival [15] [16] [17].

## 4.2 METHODS

### Patients and data

We considered a cohort of 1371 patients with Head and Neck Squamous Cell Carcinoma (HNSCC) of the oral cavity, pharynx or larynx, diagnosed at Leiden University Medical Centre. The data were obtained from files used in an earlier study [16]. The same data had been used before to derive a prediction model based on the Cox regression modelling technique [15]. Predictors in this model included Tumor location, Age at diagnosis, Gender, T-N-M classification (T=the extent of the primary tumor, N=the absence or presence and extent of regional lymph node metastasis, M=the absence or presence of distant metastasis) and Prior malignancies. In 2010, Datema et al. [16] [17] published an updated model including comorbidity according to the Adult Comorbidity Evaluation, based on a 27-item comorbidity index (ACE27) [18]. In our study, we excluded patients for whom comorbidity was unknown, resulting in a total of 1282 patients.

### Outcome variables

We defined three outcome variables related to patient survival:

- The 60-month survival (dichotomous, dead or alive, ignoring censoring before 60 months)
- The pseudo values at 60 months (continuous)
- The estimated survival time (continuous)

We focused on 60-month survival, since this is a common time point in cancer research. We subsequently calculated pseudo values for the time points 12, 24,...,288, and 300 months to reflect the individual survival patterns of patients using the R-package "Pseudo". The pseudo values form a new set of observations to allow for analysis as if we had time-to-event data without censoring [10] [11].

The estimated survival time was calculated as the sum of the pseudo values at these time points, because this sum reflects the area under the survival curve and can be interpreted as the mean survival time. The choice for a time interval of 12 months was motivated by the wish to have around 25 time intervals per subject for sufficient accuracy in estimating the survival time. Appendix 1 gives a more detailed description of the calculation and interpretation of the pseudo values and the estimated survival time. For univariate analysis of 60-month survival and overall survival we used Kaplan-Meier analysis and Cox regression analysis.

## Modelling techniques

We considered the following modelling techniques: support vector machines (SVM), neural networks (NNET), recursive partitioning (RPART), general linear models (GLM) and logistic regression (LR), with their implementations as available in the software package R, version 2.14.1 [19]. The parameters of the various modelling techniques are presented in Table 1.

**Table 1** Parameters required for the modelling techniques

| Modelling technique | Parameters |
|---|---|
| NNET | size and decay |
| RPART | cp-value |
| SVM LINEAR | cost and gamma |
| SVM POLYNOMIAL | cost, gamma and degree |
| SVM RADIAL | cost and gamma |

Appendix 2 presents a more detailed description of the various modelling techniques and their parameters, based on previous literature [20][21][22][23][24][25][26][27].

## Tuning of the modelling techniques

Before applying a modelling technique, we tuned that technique by varying the parameters to create an optimal model fit. The optimal parameter setting was based on the smallest prediction error after 10-fold cross validation. The modelling technique SVM was tuned using a simultaneous grid search for the parameters cost and gamma when a radial or linear kernel was used and for the parameters cost, gamma and degree when a polynomial kernel was used. The modelling technique NNET was tuned using a simultaneous grid search for the parameter size, and the modelling technique RPART was tuned by varying the cp-value.

## Validation and performance of the modelling techniques

For all models, internal validation was done by bootstrap resampling (200 bootstrap samples). From the original data set a bootstrap sample was drawn (randomly and with replacement). Then the modelling technique was tuned to create an optimal model fit for this bootstrap sample. With the optimal setting resulting from the tuning, we applied the modelling technique to the bootstrap sample and calculated the performance of the resulting model (bootstrap performance). We then applied the model to the original data base and calculated the performance (validated performance). This process was repeated 200 times. The 200 results were averaged to produce a single estimation of the bootstrap performance and the validated performance [28]. The dif-

ference of the mean bootstrap performance and the mean validated performance indicated the optimism of a model. The optimism corrected performance was calculated by subtracting the optimism from the apparent performance estimate, i.e. when the model was optimized and assessed for its performance on the original data set. With respect to dichotomous 60-month survival, the performance measure was the area under the ROC-curve (AUC). With respect to continuous pseudo values at 60 months and estimated survival time, the performance of the models was calculated using the root of the mean squared error (RSME).

### Variable importance

We calculated the relative importance of each of the eight predictor variables in a model by calculating the difference between the validated performance of the full model with all eight predictor variables and the validated performance of the model with seven predictor variables, leaving out each predictor variable in turn.

### Ethics statement

Patient data were used that had been collected prospectively and anonymously between 1981-1998. According to Dutch regulations, neither medical nor ethical approval was required to conduct the study, as no interventions were initiated and the study had no influence on medical care nor on decision making. The data was anonymized. The study was not supported financially in any way.

## 4.3 RESULTS

### Patients and data

Of the 1371 patients included originally, we dropped 89 patients for whom the comorbidity was unknown. As a result, we included 1282 patients in our analysis. Of these, 986 patients died during a median follow-up of 66 months (60-month survival: 52% [95% CI: 50%-55%], Figure 1). The censoring pattern of the patients (censoring rate before 60 months: 4%) is presented in Figure 2.

**Figure 1** Survival pattern 1282 patients with newly diagnosed HNSCC

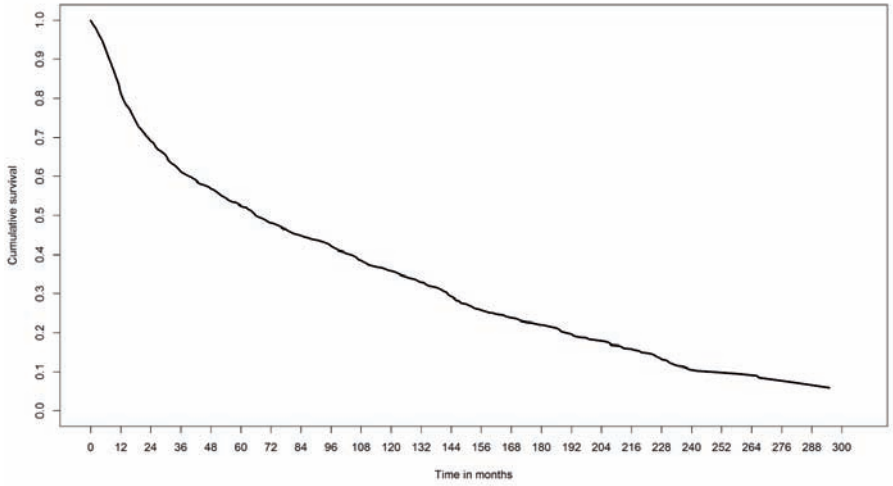



**Figure 2** Censoring pattern 1282 patients with newly diagnosed HNSCC

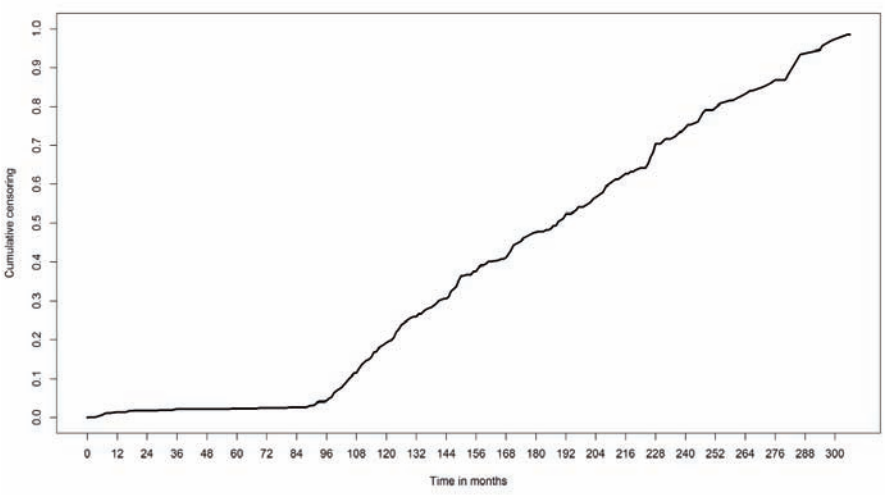



Table 2 shows the overall number of events and the survival probabilities for each category of the predictor variables with respect to the Kaplan-Meier estimated 60-month survival. Several characteristics were associated with a poor 60-month survival: Tumor location in the Hypopharynx, Oral cavity and Oropharynx (60-month survival 0.33, 0.36 and 0.37 respectively), cancer stages T3, T4, and N3 (60-month survival 0.38, 0.27, 0.11 respectively), higher age (Age >=70, 60-month survival 0.40) and severe comorbidity (Grade 3 of ACE27, 60-month survival 0.25).

**Table 2** Overall survival and 60-month survival

| | | Overall | | | | 60 months | |
|---|---|---|---|---|---|---|---|
| Variable | Value | Total (n) | Events (n) | HR | 95% CI | Survival probability | 95% CI |
| Gender | Male (ref) | 1022 | 789 | 1.00 | - | 0.54 | [0.51; 0.57] |
| | Female | 260 | 197 | 1.12 | [0.96; 1.31] | 0.48 | [0.42; 0.54] |
| Tumor location | Glottic larynx (ref) | 425 | 282 | 1.00 | - | 0.71 | [0.67; 0.75] |
| | Lip | 85 | 54 | 0.88 | [0.66; 1.18] | 0.75 | [0.67; 0.85] |
| | Oral cavity | 261 | 210 | 2.04 | [1.70; 2.44] | 0.36 | [0.31; 0.43] |
| | Oropharynx | 148 | 129 | 2.37 | [1.92; 2.92] | 0.37 | [0.30; 0.46] |
| | Nasopharynx | 39 | 23 | 1.35 | [0.88; 2.06] | 0.52 | [0.37; 0.74] |
| | Hypopharynx | 135 | 123 | 2.83 | [2.29; 3.51] | 0.33 | [0.26; 0.42] |
| | Supraglottic larynx | 189 | 165 | 1.70 | [1.40; 2.06] | 0.50 | [0.43; 0.57] |
| T-class | T1 (ref) | 454 | 293 | 1.00 | - | 0.74 | [0.70; 0.78] |
| | T2 | 354 | 281 | 1.63 | [1.38; 1.92] | 0.53 | [0.48; 0.58] |
| | T3 | 200 | 170 | 2.26 | [1.87; 2.73] | 0.38 | [0.32; 0.45] |
| | T4 | 274 | 242 | 3.18 | [2.68; 3.78] | 0.27 | [0.22; 0.33] |
| N-class | N0 (ref) | 891 | 641 | 1.00 | - | 0.64 | [0.61; 0.67] |
| | N1 | 138 | 125 | 2.10 | [1.73; 2.54] | 0.33 | [0.26; 0.42] |
| | N2 | 174 | 147 | 2.45 | [2.04; 2.94] | 0.28 | [0.22; 0.36] |
| | N3 | 79 | 73 | 3.82 | [2.99; 4.89] | 0.11 | [0.06; 0.21] |
| M-class | M0 (ref) | 1266 | 972 | 1.00 | - | 0.53 | [0.50; 0.56] |
| | M1 | 16 | 14 | 8.51 | [4.97; 14.58] | 0.00 | - |
| Prior malignancies | No (ref) | 1160 | 880 | 1.00 | - | 0.54 | [0.51; 0.57] |
| | Yes | 122 | 106 | 1.62 | [1.32; 1.98] | 0.36 | [0.28; 0.45] |
| ACE27 | Grade 0 (ref) | 782 | 574 | 1.00 | - | 0.57 | [0.54; 0.61] |
| | Grade 1 | 239 | 176 | 1.17 | [0.99; 1.39] | 0.52 | [0.46; 0.59] |
| | Grade 2 | 185 | 164 | 1.66 | [1.40; 1.98] | 0.44 | [0.38; 0.52] |
| | Grade 3 | 76 | 72 | 2.52 | [1.97; 3.23] | 0.25 | [0.17; 0.37] |
| Age class | <50 (ref) | 173 | 100 | 1.00 | - | 0.66 | [0.59; 0.74] |
| | 50-59 | 339 | 234 | 1.24 | [0.98; 1.57] | 0.59 | [0.54; 0.65] |
| | 60-69 | 404 | 328 | 1.73 | [1.38; 2.16] | 0.52 | [0.47; 0.57] |
| | >=70 | 366 | 324 | 2.53 | [2.02; 3.18] | 0.40 | [0.36; 0.46] |
| Total | | 1282 | 986 | | | 0.52 | [0.50; 0.55] |

HR: Hazard ratio
CI: Confidence interval

## Model performance and optimism

We evaluated the performance of the various models with respect to the three survival related outcome variables.

For the outcome 'dead or alive at 60 months', the LR model had the highest optimism corrected AUC (0.791, Table 3) followed by the SVM model with linear kernel (AUC 0.787, Table 3). The NNET model performed slightly poorer (AUC 0.785, Table 3). The RPART model had the lowest AUC (0.725, Table 3).

Considering the outcome 'pseudo values at 60 months', the GLM model had the highest optimism corrected RMSE (0.436, Table 4). The SVM model with polynomial kernel and the NNET model performed poorly (RMSE 0.482 and 0.486 respectively, Table 4).

Analyzing the outcome 'estimated survival time', the GLM model had the lowest optimism corrected RMSE (77.7, Table 5), followed by the SVM model with a linear kernel (79.2, Table 5). The NNET model had the worst RMSE (83.7, Table 5).

The regression based models (LR and GLM) had relatively small optimism. This small optimism was also noted for the SVM models with a linear kernel. The bootstrap-estimated optimism was substantial for NNET and the more complex SVM models with polynomial and radial kernels (Table 3 to Table 5).

**Table 3** Performance of models for the outcome 'dead or alive at 60 months'

| **Dead or alive at 60 months** | | | | | |
|---|---|---|---|---|---|
| Modelling technique | AUC bootstrap | AUC validated | AUC-apparent | Optimism | Optimism-corrected-AUC |
| LR | 0.809 | 0.797 | 0.803 | 0.012 | 0.791 |
| NNET | 0.880 | 0.810 | 0.855 | 0.070 | 0.785 |
| RPART | 0.769 | 0.741 | 0.753 | 0.028 | 0.725 |
| SVM LINEAR | 0.807 | 0.794 | 0.800 | 0.013 | 0.787 |
| SVM POLYNOMIAL | 0.861 | 0.811 | 0.821 | 0.050 | 0.771 |
| SVM RADIAL | 0.872 | 0.813 | 0.825 | 0.059 | 0.766 |

**Table 4** Performance of models for the outcome 'pseudo values at 60 months'

| **Pseudo values at 60 months** | | | | | |
|---|---|---|---|---|---|
| Modelling technique | RMSE bootstrap | RMSE validated | RMSE-apparent | Optimism | Optimism-corrected-RMSE |
| GLM | 0.427 | 0.433 | 0.430 | 0.006 | 0.436 |
| NNET | 0.388 | 0.457 | 0.417 | 0.069 | 0.486 |
| RPART | 0.430 | 0.448 | 0.448 | 0.018 | 0.466 |
| SVM LINEAR | 0.461 | 0.470 | 0.460 | 0.009 | 0.469 |
| SVM POLYNOMIAL | 0.409 | 0.445 | 0.446 | 0.036 | 0.482 |
| SVM RADIAL | 0.428 | 0.446 | 0.442 | 0.018 | 0.460 |

**Table 5** Performance of models for the outcome 'estimated survival time'

| **Estimated survival time** | | | | | |
|---|---|---|---|---|---|
| Modelling technique | RMSE bootstrap | RMSE validated | RMSE-apparent | Optimism | Optimism-corrected-RMSE |
| GLM | 76.0 | 77.1 | 76.6 | 1.1 | 77.7 |
| NNET | 80.3 | 83.0 | 81.0 | 2.7 | 83.7 |
| RPART | 76.7 | 80.1 | 79.8 | 3.4 | 83.1 |
| SVM LINEAR | 77.4 | 78.7 | 77.9 | 1.3 | 79.2 |
| SVM POLYNOMIAL | 69.7 | 76.3 | 76.3 | 6.6 | 82.9 |
| SVM RADIAL | 69.7 | 76.4 | 76.8 | 6.7 | 83.4 |

## Variable importance

For each model and for each outcome we calculated the variable importance (Figure 3). We chose the parameter settings of the modelling techniques based on the highest frequency (mode) resulting from the bootstrap procedure (Table 6).

**Table 6** Mode of the parameter settings identified as optimal in bootstrap samples

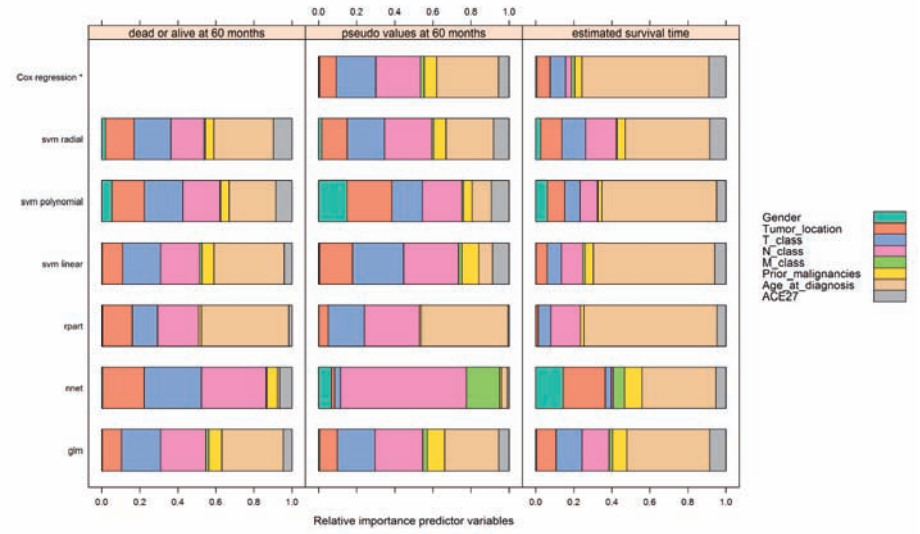
| **Outcome** | | | |
|---|---|---|---|
| Modelling technique | Dead or alive at 60 months | Pseudo values at 60 months | Estimated survival time |
| LR | - | - | - |
| NNET | size=40 | size=30 | size=40 |
| RPART | cp=0.01 | cp=0.01 | cp=0.01 |
| SVM LINEAR | cost=0.5, gamma=0.001 | cost=0.5, gamma=0.001 | cost=0.5, gamma=0.001 |
| SVM POLYNOMIAL | cost=50, gamma=0.05, degree=3 | cost=25, gamma=0.05, degree=3 | cost=50, gamma=0.05, degree=3 |
| SVM RADIAL | cost=50, gamma=0.05 | cost=0.5, gamma=0.05 | cost=50, gamma=0.05 |

Figure 3 shows the variable importance for each model and for each outcome with these parameter settings.

Overall, the variables Tumor location, T-class and N-class were the most important predictor variables for predicting the dichotomous and continuous 60-month survival (Figure 3). Survival probability was considerably lower for patients with cancer stages T4 and N3 (Appendix 3, Table 7, Table 8).

For the estimated survival time, age at diagnosis was the most important predictor variable (Figure 3). Cancer stages T1 and N0 indicated a relatively good survival prob-

**Figure 3** Variable importance of the models per outcome




*Cox regression was added as reference technique

ability (Appendix 3, Table 9). The relative importance of each predictor variable varied substantially by the specific aspect of survival studied and modelling technique used. The variable plots with observed 60-month survival (dichotomous) proved to be very similar to the variable plots with pseudo values at 60 months (continuous), except for the NNET model (Figure 3).

## 4.4 DISCUSSION

In this study, we demonstrated that pseudo values as described by Klein et al. [10] [11] enable statistically appropriate analyses of survival outcomes when used in in three variants of support vector machines (SVM), neural networks (NNET), general linear models (GLM), recursive partitioning (RPART) and logistic regression (LR). We showed that pseudo values enabled us to apply these techniques to predict survival in a case study of 1282 Dutch patients with newly diagnosed HNSCC, and to compare the performance of the resulting models.

Our analysis showed that conventional regression analysis approaches (logistic regression and the generalized linear model) outperformed the performance of relatively modern modelling techniques. However, the SVM model with an optimal setting and a linear kernel performed only slightly worse with respect to our outcomes. The NNET model and the RPART model performed relatively poorly.

We compared the performance of the alternative modelling techniques in predicting three variants of survival outcome for our case study . The first, admittedly rather simplistic, outcome variable was based on the 60-month survival in terms of dead or alive. This outcome may produce bias unless the censoring rate is small (4% in our study). The other two outcome variables were defined by means of pseudo values, which were derived from the Kaplan Meier survival function.

A drawback of outcome definitions for 60 months is that they only consider survival at a particular point in time rather than the full survival curve. By contrast, the approach with the estimated survival time is attractive, because it considers the full survival curve. We consider the total expected survival time the most relevant to inform patients about their prognosis and to support decision making.

In our study, SVM models with a linear kernel and optimal settings performed slightly worse than conventional regression modelling. These findings are in line with other studies that used support vector machines for analyzing survival [3] [4] [5] [6]. On the other hand, our findings also support the results of previous studies that relied on Cox regression modelling to predict the five year mortality and the overall mortality of newly diagnosed patients with HNSCC [15] [16] [17].

None of the investigated models showed a very satisfactory performance. This may possibly be explained by the low signal-to-noise ratio in our data. In 1998, Ennis et al. discussed the predictive performance of adaptive non-linear algorithms versus conventional statistical techniques. Based on their quite negative findings for the more modern algorithms, they postulated that adaptive non-linear methods may be most useful in problems with high signal-to-noise ratios, which sometimes occur in engineering and physical science. Since the signal-to-noise ratio is often quite low in medical prediction studies, they concluded that modern methods may have less to offer [24].

A limitation of this study is that the results were based on a single cohort of 1282 Dutch patients, diagnosed at a single center [16]. We had to rely on bootstrap validation to estimate the performance of alternative modelling techniques. On the other hand, the number of events was more than sufficient to allow for detailed statistical modelling with modern techniques for the relatively small set of candidate predictors.

We showed that the use of pseudo values opens new possibilities for analyzing survival problems with techniques other than conventional regression techniques. The validity of the pseudo value approach is supported by the concordance between Cox regression modelling for censored survival time and Generalized Estimating Equation modelling (GEE) using a log-minus-log link function [11]. Therefore, this approach deserves a central role in the ongoing search for improved prediction models for survival. On the other hand, our results also show that it may be hard to find modelling approaches that

are superior to conventional regression analysis in terms of performance, applicability and simplicity.
In conclusion, the use of pseudo values makes it readily possible to analyze survival time with alternative modelling techniques, to compare their performance and to search further for promising alternative modelling techniques to analyze survival time. In our case study on patients with newly diagnosed HNSCC, none of the alternative modelling techniques provided better predictions for survival than conventional regression modelling techniques. The estimated importance of predictors depends on the specific aspect of survival studied and the modelling technique used.

## 4.5 ACKNOWLEDGEMENTS

J. Molenaar of ONCDOC in the Leiden University Medical Centre kindly provided the data. This study is part of an ongoing research project in collaboration with the department of Otorhinolaryngology and Head and Neck Surgery, Leiden University Medical Center, Leiden, The Netherlands. The dataset is available on request for research purposes.

## 4.6 APPENDIX1

This appendix describes the calculation of the pseudo values, the calculation of the estimated survival time and the interpretation of the pseudo values [10].

### Calculation of the pseudo values and the estimated survival time

Let S(t) be the estimated Kaplan-Meier survival function. We calculated the pseudo values $J_i(t)$ for the $i^{th}$ patient as $J_i(t)=nS(t)-(n-1)S^{(-i)}(t)$ with $S^{(-i)}(t)$ the survival function without the $i^{th}$ patient.

$J_i(t)$ can be considered as the individual survival function for the $i^{th}$ patient. The area under the survival curve of $J_i(t)$ is the survival time for the $i^{th}$ patient.

In our study, we calculated the pseudo values at the time points t=12, 24,...,300. For the $i^{th}$ patient, we calculated the estimated survival time ($EST_i$) as $EST_i=12(J_i(12)+J_i(24)+...+J_i(300))$.

### Examples

The following two examples are meant as an illustration how to interpret the pseudo values of a patient. We do not consider the case of negative pseudo values.

*Example 1 (censored case):*

Suppose a patient has the following series of pseudo values at the particular time points:

| Time point | 12 | 24 | 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 | 132 | 144 | 156 | 168 | 180 | 192 | 204 | 216 | 228 | 240 | 252 | 264 | 276 | 288 | 300 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pseudo value | 1.00 | 1.00 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.02 | 1.03 | 1.05 | 1.09 | 1.14 | 1.18 | 1.23 | 1.31 | 1.35 | 1.19 | 0.98 | 0.80 | 0.74 | 0.71 | 0.64 | 0.55 | 0.44 |

The interpretation of the series of pseudo values for this patient is:

The patient did not die because all pseudo values are positive. The follow-up time point lies between the time points 204 and 216 because the increasing pattern of the pseudo values changes into a decreasing pattern.

*Example 2 (uncensored case):*

Suppose a patient has the following series of pseudo values at the particular time points:

| Time point | 12 | 24 | 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 | 132 | 144 | 156 | 168 | 180 | 192 | 204 | 216 | 228 | 240 | 252 | 264 | 276 | 288 | 300 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pseudo value | 1.00 | 1.00 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.02 | -0.19 | -0.18 | -0.16 | -0.14 | -0.13 | -0.12 | -0.10 | -0.10 | -0.08 | -0.07 | -0.06 | -0.05 | -0.05 | -0.05 | -0.04 | -0.03 |

The interpretation of the series of pseudo values for this patient is:

The patient died between the time points 108 and 120 because the pseudo values change from a positive value into a negative value.

Figure 4 shows that the Kaplan-Meier survival curve nearly matches the curve of the mean pseudo values at each time point.

Figure 5 shows that the sum of the pseudo values multiplied by 12 can be considered as an estimation of the area under the survival curve and therefore as an estimation of the expected survival time for a patient.
For the modelling of the various models we therefore used this variable as outcome variable to estimate the survival time of a patient.

**Figure 4** Comparison KM survival curve and survival curve based on pseudo values

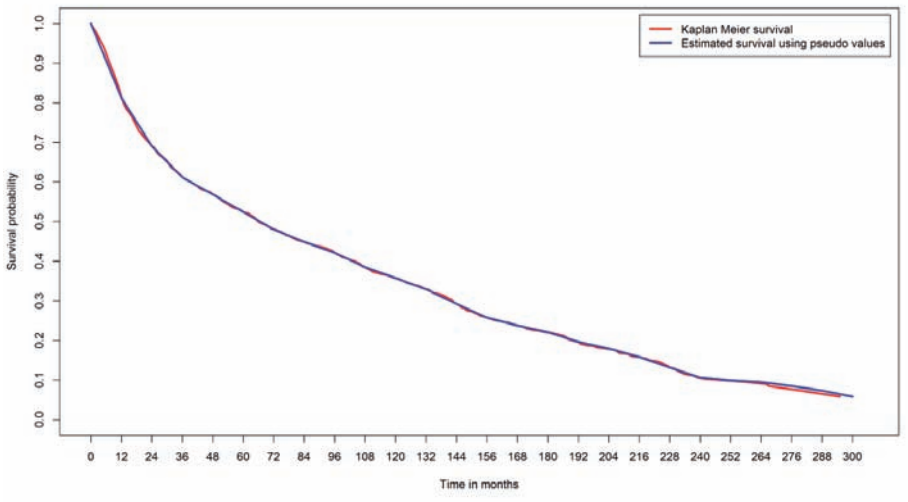
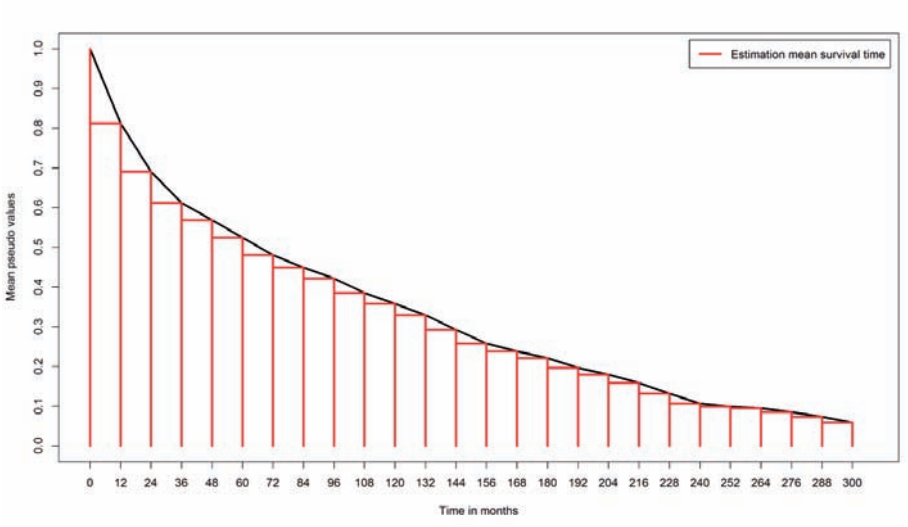



**Figure 5** Estimation of the area under the survival curve using pseudo values



## 4.7 APPENDIX 2

This appendix contains a description of the evaluated modelling techniques and their parameters, based on literature and articles of various authors [20-24]. The standard regression techniques logistic regression (LR) and generalized linear model (GLM) are assumed to be familiar.

### Support vector machines

A Support Vector Machine (SVM) performs classification tasks by constructing hyperplanes with a margin in a multidimensional space that separates cases from different classes. SVM can efficiently perform a non-linear classification or regression task using different kernels (radial, linear and polynomial). The tuning parameters for SVM are the C-parameter (cost), which regulates the margin width, and the gamma-parameter for the kernel calculation. SVM claims to be a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM may particularly be suited to analyze data with large numbers of predictor variables.

### Neural net

A neural network (NNET), sometimes called a multilayer perceptron, works by simulating a large number of interconnected simple processing units, which are arranged in layers. There are three parts in a neural network: an input layer, with units representing the predictor variables, one or more hidden layers and an output layer, with a unit representing the outcome variable. The units are connected with varying connection strengths or weights. Input data are presented to the input layer and values are propagated from there to the next layer. Then, a prediction is delivered from the output layer. The network learns by examining individual records, generating a prediction for each record and making adjustments to the weights whenever it makes an incorrect prediction. This process is repeated many times, and the network continues to improve its predictions until one or more of the stopping criteria have been met. Initially, all weights are random, and the predictions that come out of the net are nonsensical. The network learns through training. Records for which the output is known are repeatedly presented to the network, and the predictions it gives are compared to the known outcomes. As training progresses, the network becomes increasingly accurate in replicating the known outcomes. Once trained, the network can be applied to new patients for whom the outcome is unknown. The parameters of NNET are the size-parameter (number of units in the layer) and decay-parameter.

## Recursive partitioning

Recursive partitioning (RPART) is a tree-based classification and prediction modelling technique which uses recursive partitioning to split the training records into segments with similar output variable values. The modelling starts by examining the input variables to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two further subgroups and so on, until the stopping criterion is met. The parameter of RPART is the cp-parameter (cost complexity factor). A cp-value of 0.001 for example regulates that a split must decrease the overall lack of fit by a factor of 0.001.

## 4.8 APPENDIX 3

This appendix contains the coefficients of the regression models.

### Coefficients regression models

**Table 7** Logistic regression model for the outcome 'dead or alive at 60 months'

| Variable | Value | B | SE | P-value | OR | 95% CI |
|---|---|---|---|---|---|---|
| Tumor location | Glottic larynx (ref) | 0.00 | - | - | 1.00 | - |
| | Lip | 0.04 | 0.31 | 0.89 | 1.05 | [0.57; 1.91] |
| | Oral cavity | 1.00 | 0.21 | 0.00 | 2.73 | [1.83; 4.07] |
| | Oropharynx | 0.76 | 0.25 | 0.00 | 2.15 | [1.32; 3.50] |
| | Nasopharynx | −0.09 | 0.41 | 0.82 | 0.91 | [0.41; 2.03] |
| | Hypopharynx | 0.80 | 0.26 | 0.00 | 2.21 | [1.33; 3.68] |
| | Supraglottic larynx | 0.39 | 0.22 | 0.07 | 1.48 | [0.97; 2.26] |
| ACE27 | Grade 0 (ref) | 0.00 | - | - | 1.00 | - |
| | Grade 1 | 0.04 | 0.18 | 0.82 | 1.04 | [0.74; 1.47] |
| | Grade 2 | 0.36 | 0.19 | 0.06 | 1.43 | [0.99; 2.08] |
| | Grade 3 | 1.09 | 0.31 | 0.00 | 2.97 | [1.62; 5.45] |
| T-class | T1 (ref) | 0.00 | - | - | 1.00 | - |
| | T2 | 0.67 | 0.17 | 0.00 | 1.95 | [1.38; 2.74] |
| | T3 | 0.90 | 0.21 | 0.00 | 2.47 | [1.62; 3.76] |
| | T4 | 1.30 | 0.21 | 0.00 | 3.68 | [2.44; 5.55] |
| N-class | N0 (ref) | 0.00 | - | - | 1.00 | - |
| | N1 | 0.73 | 0.22 | 0.00 | 2.08 | [1.34; 3.22] |
| | N2 | 1.02 | 0.22 | 0.00 | 2.76 | [1.81; 4.22] |
| | N3 | 2.13 | 0.38 | 0.00 | 8.40 | [3.98; 17.72] |
| M-class | M0 (ref) | 0.00 | - | - | 1.00 | - |
| | M1 | 1.65 | 0.85 | 0.05 | 5.23 | [0.99; 27.63] |
| Prior malignancies | No (ref) | 0.00 | - | - | 1.00 | - |
| | Yes | 1.04 | 0.24 | 0.00 | 2.83 | [1.78; 4.50] |
| Gender | Male (ref) | 0.00 | - | - | 1.00 | - |
| | Female | −0.05 | 0.17 | 0.77 | 0.95 | [0.68; 1.33] |
| Age at diagnosis per decade | | 0.49 | 0.06 | 0.00 | 1.63 | [1.44; 1.84] |
| Constant | | −4.79 | 0.44 | 0.00 | 0.01 | - |

B: Regression coefficient
SE: Standard error regression coefficient
OR: Odds ratio
CI: Confidence interval

**Table 8** General linear model for the outcome 'pseudo values at 60 months'

| Variable | Value | B | SE | 95% CI | P-value |
|---|---|---|---|---|---|
| Tumor location | Glottic larynx (ref) | 0.00 | - | - | - |
| | Lip | −0.00 | 0.05 | [−0.11; 0.10] | 0.93 |
| | Oral cavity | −0.19 | 0.04 | [−0.26; −0.11] | 0.00 |
| | Oropharynx | −0.14 | 0.05 | [−0.23; −0.05] | 0.00 |
| | Nasopharynx | −0.06 | 0.08 | [−0.21; 0.09] | 0.44 |
| | Hypopharynx | −0.15 | 0.05 | [−0.25; −0.06] | 0.00 |
| | Supraglottic larynx | −0.07 | 0.04 | [−0.15; 0.01] | 0.08 |
| ACE27 | Grade 0 (ref) | 0.00 | - | - | - |
| | Grade 1 | 0.00 | 0.03 | [−0.06; 0.06] | 0.99 |
| | Grade 2 | −0.07 | 0.04 | [−0.14; 0.00] | 0.06 |
| | Grade 3 | −0.19 | 0.05 | [−0.29; −0.09] | 0.00 |
| T-class | T1 (ref) | 0.00 | - | - | - |
| | T2 | −0.13 | 0.03 | [−0.20; −0.07] | 0.00 |
| | T3 | −0.19 | 0.04 | [−0.27; −0.11] | 0.00 |
| | T4 | −0.27 | 0.04 | [−0.34; −0.19] | 0.00 |
| N-class | N0 (ref) | 0.00 | - | - | - |
| | N1 | −0.16 | 0.04 | [−0.25; −0.08] | 0.00 |
| | N2 | −0.22 | 0.04 | [−0.29; −0.14] | 0.00 |
| | N3 | −0.37 | 0.05 | [−0.47; −0.26] | 0.00 |
| M-class | M0 (ref) | 0.00 | - | - | - |
| | M1 | −0.27 | 0.11 | [−0.49; −0.05] | 0.02 |
| Prior malignancies | No (ref) | 0.00 | - | - | - |
| | Yes | −0.20 | 0.04 | [−0.28; −0.12] | 0.00 |
| Gender | Male (ref) | 0.00 | - | - | - |
| | Female | 0.01 | 0.03 | [−0.05; 0.07] | 0.69 |
| Age at diagnosis per decade | | −0.09 | 0.01 | [−0.11; −0.07] | 0.00 |
| Constant | | 1.38 | 0.07 | [1.24; 1.52] | 0.00 |

B: Regression coefficient
SE: Standard error regression coefficient
CI: Confidence interval

**Table 9** General linear model for the outcome 'estimated survival time'

| Variable | Value | B | SE | 95% CI | P-value |
|---|---|---|---|---|---|
| Tumor location | Glottic larynx (ref) | 0.00 | - | - | - |
| | Lip | −0.79 | 9.44 | [−19.30; 17.72] | 0.93 |
| | Oral cavity | −31.29 | 6.79 | [−44.59; −17.98] | 0.00 |
| | Oropharynx | −38.62 | 8.26 | [−54.82; −22.42] | 0.00 |
| | Nasopharynx | −21.39 | 13.66 | [−48.17; 5.38] | 0.12 |
| | Hypopharynx | −44.97 | 8.59 | [−61.81; −28.13] | 0.00 |
| | Supraglottic larynx | −23.41 | 7.23 | [−37.59; −9.24] | 0.00 |
| ACE27 | Grade 0 (ref) | 0.00 | - | - | - |
| | Grade 1 | −2.43 | 5.75 | [−13.69; 8.83] | 0.67 |
| | Grade 2 | −24.39 | 6.41 | [−36.95; −11.83] | 0.00 |
| | Grade 3 | −41.36 | 9.37 | [−59.72; −23.01] | 0.00 |
| T-class | T1 (ref) | 0.00 | - | - | - |
| | T2 | −25.71 | 5.79 | [−37.06; −14.35] | 0.00 |
| | T3 | −30.65 | 7.18 | [−44.72; −16.58] | 0.00 |
| | T4 | −46.44 | 6.93 | [−60.02; −32.86] | 0.00 |
| | N0 (ref) | 0.00 | - | - | - |
| N-class | N1 | −27.65 | 7.60 | [−42.54; −12.76] | 0.00 |
| | N2 | −36.42 | 7.16 | [−50.45; −22.40] | 0.00 |
| | N3 | −56.29 | 9.70 | [−75.29; −37.28] | 0.00 |
| M-class | M0 (ref) | 0.00 | - | - | - |
| | M1 | −47.31 | 19.71 | [−85.94; −8.68] | 0.02 |
| Prior malignancies | No (ref) | 0.00 | - | - | - |
| | Yes | −38.03 | 7.52 | [−52.76; −23.29] | 0.00 |
| Gender | Male (ref) | 0.00 | - | - | - |
| | Female | 2.99 | 5.56 | [−7.91; 13.90] | 0.59 |
| Age at diagnosis per decade | | −22.71 | 1.88 | [−26.39; −19.03] | 0.00 |
| Constant | | 300.47 | 12.58 | [275.82; 325.12] | 0.00 |

B: Regression coefficient
SE: Standard error regression coefficient
CI: Confidence interval

## REFERENCE LIST


1. Kaplan EL, Meier P (1958) Nonparametric Estimation from Incomplete Observations. J Am Stat Assoc 53: 457–481. Available: http://www.jstor.org/stable/2281868\nhttp://www.jstor.org/stable/pdfplus/2281868.pdf?acceptTC=true.
2. Cox DR (1972) Regression Models and Life-Tables. J R Stat Soc Ser B 34: 187–220. doi: 10.2307/2985181.
3. Mangasarian Y-JLOL, Wolberg WH (2000) Breast cancer survival and chemotherapy: a support vector machine analysis. Discrete Mathematical Problems with Medical Applications: DIMACS Workshop Discrete Mathematical Problems with Medical Applications, December 8-10, 1999, DIMACS Center. Vol. 55. p. 1.
4. Chen S, Härdle WK, Moro RA (2006) Estimation of default probabilities with support vector machines.
5. Intrator O, Kooperberg C (1995) Trees and splines in survival analysis. Stat Methods Med Res 4: 237–261. doi:10.1177/096228029500400305.
6. Biganzoli E, Boracchi P, Mariani L, Marubini E (1998) Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Stat Med 17: 1169–1186. doi:10.1002/(SICI)1097-0258(19980530)17:10<1169::AID-SIM796>3.0.CO;2-D.
7. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Ann Appl Stat: 841–860.
8. Ture M, Tokatli F, Kurt Omurlu I (2009) The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data. Expert Syst Appl 36: 8247–8254. doi:10.1016/j.eswa.2008.10.014.
9. Schwarzer G, Vach W, Schumacher M (2000) On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Statistics in Medicine. Vol. 19. pp. 541–561. doi:10.1002/(SICI)1097-0258(20000229)19:4<541::AID-SIM355>3.0.CO;2-V.
10. Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP (2008) SAS and R functions to compute pseudo-values for censored data regression. Comput Methods Programs Biomed 89: 289–300.
11. Andersen PK, Hansen MG, Klein JP (2004) Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime Data Analysis. Vol. 10. pp. 335–350. doi: 10.1007/s10985-004-4771-0.
12. Van Der Schroeff MP, Van De Schans SAM, Piccirillo JF, Langeveld TPM, Baatenburg De Jong RJ, et al. (2010) Conditional relative survival in head and neck squamous cell carcinoma: Permanent excess mortality risk for long-term survivors. Head Neck 32: 1613–1618. doi: 10.1002/hed.21369.
13. Van Der Schroeff MP, Derks W, Hordijk GJ, De Leeuw RJ (2007) The effect of age on survival and quality of life in elderly head and neck cancer patients: A long-term prospective study. Eur Arch Oto-Rhino-Laryngology 264: 415–422. doi:10.1007/s00405-006-0203-y.

14. Yung KC, Piccirillo JF (2008) The incidence and impact of comorbidity diagnosed after the onset of head and neck cancer. Arch Otolaryngol Head Neck Surg 134: 1045–1049. doi: 10.1001/archotol.134.10.1045.

15. Baatenburg de Jong RJ, Hermans J, Molenaar J, Briaire JJ, le Cessie S (2001) Prediction of survival in patients with head and neck cancer. Head Neck 23: 718–724. Available: http://www.ncbi.nlm.nih.gov/pubmed/12001073.

16. Datema FR, Ferrier MB, Van Der Schroeff MP, Baatenburg De Jong RJ (2010) Impact of comorbidity on short-term mortality and overall survival of head and neck cancer patients. Head Neck 32: 728–736. doi:10.1002/hed.21245.

17. Datema FR, Moya A, Krause P, Bäck T, Willmes L, et al. (2012) Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression. Head Neck 34: 50–58. doi:10.1002/hed.21698.

18. Piccirillo JF (2000) Importance of comorbidity in head and neck cancer. Laryngoscope 110: 593–602. doi:10.1097/00005537-200004000-00011.

19. R Development Core Team R (2011) R: A Language and Environment for Statistical Computing. R Found Stat Comput 1: 409. Available: http://www.r-project.org.

20. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees.

21. Steyerberg EW (2009) Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Available: http://www.amazon.com/Clinical-Prediction-Models-Development-Validation/dp/1441926488/ref=sr_1_1?ie=UTF8&s=books&qid=1267137425&sr=1-1.

22. Zani S (2006) Data analysis, classification and the forward search: proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6-8, 2005. Springer.

23. Harrell FE (2001) Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. Springer.

24. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R (1998) A comparison of statistical learning methods on the Gusto database. Stat Med 17: 2501–2508.

25. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF (2000) Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small data sets. Stat Med 19: 1059–1079. doi:10.1002/(SICI)1097-0258(20000430)19:8<1059::AID-SIM412>3.0.CO;2-0.

26. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44: 837–845. doi:10.2307/2531595.

27. Tufféry S (2011) Data mining and statistics for decision making. John Wiley & Sons.

# Chapter 5

## Feature selection and validated predicted performance in the domain of legionella pneumophila: a comparative study

Tjeerd van der Ploeg, Ewout W. Steyerberg

BMC Res Notes. 2016 Mar 8;9(1):147. doi: 10.1186/s13104-016-1945-2.

## ABSTRACT

### Introduction

Genetic comparisons of clinical and environmental *Legionella* strains form an essential part of outbreak investigations. DNA microarrays often comprise many DNA markers (features). Feature selection and the development of prediction models are particularly challenging in this domain with many variables and comparatively few subjects or data points. We aimed to compare modelling strategies to develop prediction models for classifying infections as clinical or environmental.

### Methods

We applied a bootstrap strategy for preselecting important features to a database containing 222 Legionella pneumophila strains with 448 continuous markers and a dichotomous outcome (clinical or environmental). Feature selection was done with 50 bootstrap samples resulting in a top 10 of most important features for each of four modelling techniques: classification and regression trees (CART), random forests (RF), support vector machines (SVM) and least absolute shrinkage and selection operator (LASSO). Validation was done in a second bootstrap re-sampling loop (200x) for evaluation of discriminatory model performance according to the AUC.

### Results

The top 5 of selected features differed considerably between the various modelling techniques, with only one common feature ("LePn.007B8"). The mean validated AUC-values of the SVM model and the CART model were 0.859 and 0.873 respectively. The LASSO and the RF model showed higher validated AUC-values (0.925 and 0.975 respectively).

### Conclusions

In the domain of Legionella pneumophila, which comprises many potential features for classifying of infections as clinical or environmental, the RF and LASSO techniques provide good prediction models. The identification of potentially biologically relevant features is highly dependent on the technique used, and should hence be interpreted with caution.

## 5.1 INTRODUCTION

The bacterium Legionella pneumophila, the causative agent for Legionnaires' disease, is omnipresent in both natural and man-made aquatic environments. The major route of transmission is inhalation of the bacterium, which is spread into the air as an aerosol from its reservoir [1]. Genetic comparisons of clinical and environmental Legionella strains form an essential part of outbreak investigations [2] [3]. Such investigations previously showed that the distribution of genotypes within clinical strains significantly differed from the distribution in environmental strains [4] [5] [6].

To develop reliable statistical models for the discrimination between clinical and environmental strains, modelling techniques are required which can stabilize the feature selection. DNA microarrays may comprise thousands of DNA markers (features, *p*) and only a few hundred or even only a few dozen subjects (*n*; the "p>n" problem) [7].

Common statistical approaches for selecting features include filter methods, wrapper methods and embedded methods. Filter methods preselect features using a univariate technique with respect to the outcome (T-test, Mann-Whitney-test, Pearson correlation coefficients), without being tuned to a specific type of modelling technique. By contrast, wrapper methods use a specific modelling technique to select features, and subsequently each selected feature set is used to train a model built with that same modelling technique; the performance of the model is usually tested on a hold-out set, resulting in a score for a specific feature set. Embedded methods are a catch-all group of techniques that perform feature selection as part of the model construction process [8] [9].

Popular feature selection methods nowadays are the least absolute shrinkage and selection operator method (LASSO) [10], recursive feature elimination, which is commonly used with support vector machines (SVM RFE) [11], and a backward feature selection method based on random forests (VARSEL RF) [12]. For stabilizing the feature selection, several authors proposed the use of ensemble feature selection based on bootstrap samples [13] [14] [15], a widely used technique in prediction research [16]. Several authors discussed double bootstrap or nested bootstrap procedures for both feature selection and performance estimation [17] [18] [19] [20] [21] [22].

The aim of the present study was to compare statistical models that can be used to discriminate between clinical and environmental strains using a small number of features. We compared modelling techniques for developing prediction models with relevant genomic features related to pathogenity. We focused on four modelling techniques: classification and regression trees (CART) [23], random forests (RF) [24], support vector machines (SVM) [25] and least absolute shrinkage and selection operator (LASSO) [26]. We used a nested bootstrap procedure, one for feature selection and one for predictive

performance validation for a fair evaluation of a prediction model based on a relatively small data set.

## 5.2 METHODS

### Data

We analyzed the database of the Dutch National Legionella Outbreak Detection Programme as used before [27]. The database contained 222 Legionella pneumophila strains with 448 continuous markers and a dichotomous outcome. Of these strains, 49 were patient-derived strains from notified cases in the Netherlands in the period 2002–2006, and 173 were environmental strains that were collected during the source investigation for those patients. The 448 continuous markers were coded as LePn.###L## (e.g. LePn.032E12). The data were collected prospectively and anonymously. According to Dutch regulations, neither medical nor ethical approval was required to conduct the study, as no medical interventions were initiated and the study had no influence on medical care nor on decision making.

### Modelling techniques

We evaluated the modelling techniques CART, RF, SVM and LASSO, which are described below.

#### *Classification and regression trees (CART)*

The CART model is a tree-based classification and prediction model that uses recursive partitioning to split the training records into segments with similar output variable values. The modelling starts by examining the input variables to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two further subgroups and so on, until the stopping criterion is met [23].

#### *Random forest (RF)*

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees [24].
Each tree is constructed using the following algorithm:

1. Let the number of training cases be N, and the number of variables in the classifier be M.
2. We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much lower than M.

3. Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
4. For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.
5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the mode of the votes over all trees is used as the random forest prediction.

*Support vector machine (SVM)*

A support vector machine performs classification tasks by constructing hyperplanes in a multidimensional space that separate cases from non-cases. It claims to be a robust technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM may particularly be suited to analyze data with large numbers of predictor variables. SVM has applications in many disciplines, including customer relationship management, image recognition, bioinformatics, text mining concept extraction, intrusion detection, protein structure prediction, and voice and speech recognition [25].

*Least absolute shrinkage selection operator (LASSO)*

Given a set of input measurements $x_1$, $x_2$,...,$x_p$ and an outcome measurement y, the LASSO fits a linear model: $\hat{y}=b_0+b_1*x_1+b_2*x_2+...+b_p*x_p$

It uses the following criterion: Minimize $\text{sum}((y-\hat{y})^2)$ subject to $\text{sum}(|b_j|)<=s$.

The first sum is taken over the cases in the dataset. The bound "s" is a tuning parameter. If "s" is large, the constraint has no effect and the solution is just the usual maximum likelihood regression of y on $x_1$, $x_2$,...,$x_p$. For smaller values of s (s>=0) the regression coefficients are shrunken versions of the maximum likelihood estimates. Often, some of the coefficients $b_j$ are shrunk to zero. We used cross-validation to estimate the best value for "s" [26], and a logistic link function rather than linear regression.

## Reference techniques

As reference points for this evaluation, we applied the commonly used techniques VARSEL RF and SVM RFE to our database, which are examples of embedded methods. VARSEL RF is a feature selection technique based on random forests with backward stepwise elimination of features that are not important. SVM RFE is a recursive feature

elimination technique. It is based on support vector machines, which eliminate feature redundancy resulting in compact feature sets.

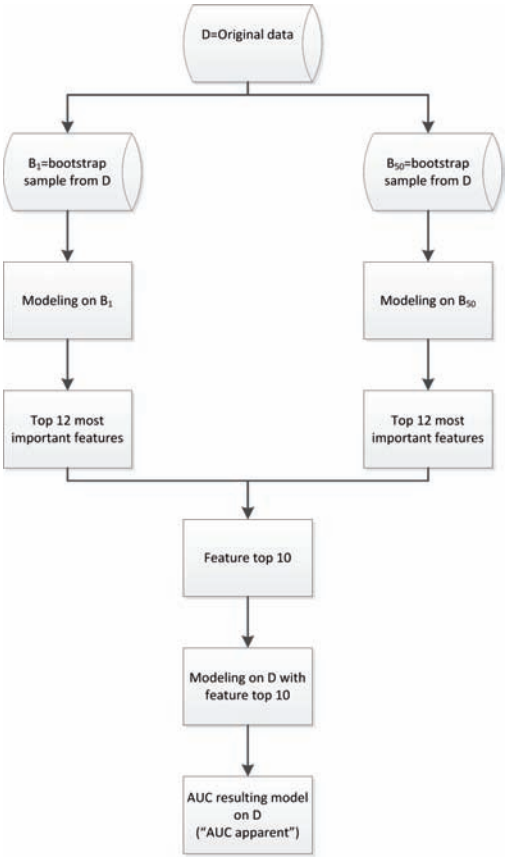
### Model performance

We evaluated the stability of the feature selection and the validated performance by means of bootstrap re-sampling from the original database. The performance of a model resulting from a modelling technique was assessed using the area under the Receiver Operator Curve (AUC).

### Modelling strategy

For a specific modelling technique, feature selection was done by bootstrap re-sampling from the original database D. We re-sampled 50 bootstraps $B_i$ ($B_1$,...,$B_{50}$) from the original database D. We applied the specific modelling technique on each $B_i$ and

**Figure 1** Feature selection and model development strategy

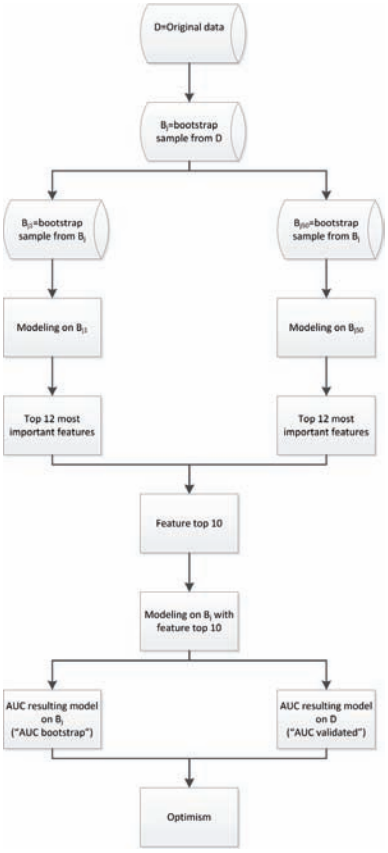


determined for each $B_i$ the top 12 of most important features, leading to 50x12=600 important features. From these 600 features, the top 10 of features with the highest frequency was extracted. With this feature top 10, a model was developed on the original database D with the specific modelling technique. For the resulting model the performance for the original database D was calculated ("AUC apparent", Figure 1)

## Validation of the strategy

To validate our strategy for a specific modelling technique, we performed a bootstrap procedure. We re-sampled a bootstrap sample $B_j$ from the original data base D and from this bootstrap sample $B_j$, we re-sampled 50 independent bootstraps $B_{ji}$ ($B_{j1},\ldots,B_{j50}$). We applied the specific modelling technique on each $B_{ji}$ and determined for each $B_{ji}$ the top 12 of most important features, leading to 50x12=600 important features. From these 600 features, the top 10 of features with the highest frequency was extracted.

**Figure 2** Evaluation of optimism for each strategy



With this top 10 features, a model was developed on bootstrap sample $B_j$ with the specific modelling technique. For the resulting model, the performance for $B_j$ and the performance for the original data base D were calculated ("AUC bootstrap" and "AUC validated" respectively). The optimism of the resulting model was calculated as "AUC bootstrap" minus "AUC validated". This process was repeated 200 times ($B_1$ to $B_{200}$, Figure 2).

### Analysis

For the modelling and the analysis of these techniques, we used R 2.14, using default settings as far as possible. We used the libraries randomForest, caTools, rpart, caret, e1071, varSelRF and glmnet [28].

## 5.3 RESULTS

### Reference techniques

Feature selection with the reference techniques VARSEL RF and SVM RFE resulted in two different sets of features, only with LePn.007B8 as the common feature in the top 5 (Table 1). For the full list of features for each technique and for each bootstrap sample, we refer to file S1 and file S2. The mean validated AUC-values of the models generated by these two techniques were 0.966 for VARSEL RF and 0.915 for SVM RFE (Table 2).

### Other techniques

The top 5 of selected features differed among the other modelling techniques (CART, RF, SVM, LASSO). The only common feature in the top 5 of all four modelling techniques was feature LePn.007B8. Feature selection with RF resulted in four matches with feature selection based on VARSEL RF, and feature selection with LASSO resulted in three matches with feature selection with SVM RFE (Table 3). The selected features also differed within the various modelling techniques. For the full list of selected features for each technique and for each bootstrap sample, we refer to file S3, file S4, file S5 and file S6. The RF model showed the highest mean validated AUC-value (0.975) followed by the LASSO model (0.925). The mean validated AUC-values of the CART and the SVM models were 0.873 and 0.859 respectively (Table 4). The RF model showed a relatively low statistical optimism (0.005). Modelling with CART, SVM and LASSO resulted in prediction models with higher optimism (decrease in performance 0.064, 0.066 and 0.056 respectively, Table 4).

**Table 1** Top 5 features VARSEL RF and SVM RFE and frequency of selection in 200 bootstrap resamples.

| Technique | Top 5 features and frequencies [ ] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VARSELRF | LePn.007B8 | [ 196 ] | LePn.032E12 | [ 93 ] | LePn.004E8 | [ 71 ] | LePn.015B2 | [ 40 ] | LePn.035C6 | [ 40 ] |
| SVMRFE | LePn.007B8 | [ 88 ] | LePn.016E4 | [ 80 ] | LePn.033H2 | [ 77 ] | LePn.005H6 | [ 60 ] | LePn.033D7 | [ 54 ] |

**Table 2** Mean AUC and mean optimism VARSEL RF and SVM RFE

| | | Bootstrap AUC | | Validated AUC | | Optimism | |
|---|---|---|---|---|---|---|---|
| **Technique** | **Apparent AUC** | **Mean** | **95% CI** | **Mean** | **95% CI** | **Mean** | **95% CI** |
| VARSELRF | 0.904 | 0.966 | [0.963;0.969] | 0.966 | [0.963;0.969] | 0.000 | [−0.004;0.004] |
| SVMRFE | 0.964 | 0.991 | [0.990;0.992] | 0.915 | [0.911;0.919] | 0.076 | [0.072;0.080] |

**Table 3** Top 5 features CART, RF, SVM and LASSO and frequency of selection in 200 bootstrap resamples.

| Technique | Top 5 features and frequencies [ ] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CART | LePn.007B8 | [ 200 ] | LePn.026A7 | [ 93 ] | LePn.027A12 | [ 76 ] | LePn.028A11 | [ 71 ] | LePn.016E4 | [ 66 ] |
| RF | LePn.007B8 | [ 200 ] | LePn.032E12 | [ 168 ] | LePn.004E8 | [ 151 ] | LePn.035C6 | [ 141 ] | LePn.016E4 | [ 100 ] |
| SVM | LePn.007B8 | [ 144 ] | LePn.035G3 | [ 111 ] | LePn.009C5 | [ 105 ] | LePn.012C5 | [ 97 ] | LePn.024C3 | [ 89 ] |
| LASSO | LePn.007B8 | [ 187 ] | LePn.033H2 | [ 146 ] | LePn.016E4 | [ 131 ] | LePn.010B12 | [ 83 ] | LePn.011B3 | [ 77 ] |

**Table 4** Mean AUC and mean optimism CART, RF, SVM and LASSO

| | | Bootstrap AUC | | Validated AUC | | Optimism | |
|---|---|---|---|---|---|---|---|
| **Technique** | **Apparent AUC** | **Mean** | **95% CI** | **Mean** | **95% CI** | **Mean** | **95% CI** |
| CART | 0.929 | 0.937 | [0.933;0.942] | 0.873 | [0.868;0.878] | 0.064 | [0.060;0.068] |
| RF | 0.938 | 0.980 | [0.978;0.981] | 0.975 | [0.973;0.976] | 0.005 | [0.003;0.008] |
| SVM | 0.887 | 0.924 | [0.918;0.930] | 0.859 | [0.852;0.866] | 0.066 | [0.061;0.071] |
| LASSO | 0.965 | 0.981 | [0.980;0.983] | 0.925 | [0.922;0.928] | 0.056 | [0.053;0.060] |

## 5.4 DISCUSSION

Using a feature selection and validation strategy based on bootstrap procedures, we found that RF and LASSO modelling resulted in prediction models with high performance. The statistical optimism of the RF model was relatively low (0.005). By contrast, modelling with CART, SVM and LASSO resulted in prediction models which had a good validated performance, but with higher optimism in the apparent performance estimates (0.064, 0.066 and 0.056 respectively).

We applied two commonly used techniques as references: variable selection from random forests using backward variable elimination (VARSEL RF) and support vector machines using recursive feature elimination (SVM RFE). We applied these techniques to the same database and validated the resulting models by means of bootstrap resampling. These analyses showed that VARSEL RF had a high validated performance (AUC 0.966), whereas modelling with SVM RFE resulted in a validated performance of 0.915 and an optimism of 0.076.

We used the bootstrap procedure as described by Efron [16]. The original data set comprised 222 Legionella strains. Bootstrapping from that data set leads to 222 Legionella strains again in each bootstrap sample because it is based on simple re-sampling with replacement. We note than the 0.632+ variant of the standard bootstrap validation procedure uses only cases not used at model development. Empirical evaluations for binary prediction showed no advantage of this bootstrapping variant [29] . Hence, we did not use this approach in the estimation of the optimism of the models and the stability of the feature set.

Our results are in line with earlier findings, which showed that RF and LASSO are suitable modelling techniques for feature selection and that the resulting models have a good predictive performance [10] [11]. Our results with SVM modelling are in line with the work of Guyon et al., who suggested SVM RFE for feature selection [11]. However, the features selected with SVM and bootstrapping differed from the features selected with the SVM RFE approach. The validated predictive performance of our strategy with SVM modelling was inferior to the validated predictive performance with the SVM RFE approach (mean validated AUC 0.859 and 0.915 respectively).

We found that feature selection by means of VARSEL RF resulted in models with a high validated performance. This is in line with the findings of earlier studies that used a simpler validation procedure [27]. Likewise, RF modelling resulted in models with a very high performance (mean validated AUC 0.975). Feature selection with either of the two RF approaches resulted in four matching features (LePn.007B8, LePn.004E8, LePn.032E12 and LePn.035C6).

Feature selection with LASSO modelling resulted in a top 3 that was identical to the top 3 based on feature selection with SVM RFE. The relevance of this match is reinforced

by the fact that feature selection with both these techniques resulted in models with a fairly high performance (validated AUC 0.915 and 0.925 respectively).
One of the limitations of our study is that we used one single database with features of a specific bacterium to compare the performance of the various modelling techniques. Future research should apply strong validation methods, such as our double bootstrap method, when analyzing comparable databases, such as databases comprising Legionella strains from other countries. An even stronger validation would be achieved by testing the models on new, independent data. Another limitation is that we restricted our research to four modelling techniques (CART, RF, SVM and LASSO). Various other techniques might also be suitable for feature selection and prediction in a domain with many variables and few subjects.

### Conclusions

In the domain of Legionella pneumophila, which comprises many potential features for classifying of infections as clinical or environmental, the RF and LASSO techniques provide good prediction models. The identification of potentially biologically relevant features is highly dependent on the technique used, and should hence be interpreted with caution.

## 5.5 ABBREVIATIONS

| | |
|---|---|
| DNA: | Deoxyribonucleic acid; |
| LASSO: | Least absolute shrinkage and selection operator; |
| SVM RFE: | Support vector machines recursive feature elimination; |
| VARSEL RF: | Variable selection random forest; |
| CART: | Classification and regression trees; |
| RF: | Random forest; |
| SVM: | Support vector machines. |

## 5.6 ACKNOWLEDGEMENTS

The authors thank Michel Ossendrijver, Frank Schuren (Department of Microbiology, Netherlands Organisation of Applied Scientific Research TNO, Zeist, The Netherlands), Jeroen den Boer and Sjoerd Euser (Regional Public Health Laboratory Kennemerland, Haarlem, The Netherlands) and Nico Nagelkerke (Erasmus Medical Center, Rotterdam, The Netherlands) for methodological and statistical advice. The authors thank Lisette van Hulst (Medical Center Alkmaar, Alkmaar, The Netherlands) for editorial support.

## 5.7 SUPPORTING FILES

S1: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782323/bin/13104_2016_1945_MOESM1_ESM.xlsx

S2: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782323/bin/13104_2016_1945_MOESM2_ESM.xlsx

S3: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782323/bin/13104_2016_1945_MOESM3_ESM.xlsx

S4: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782323/bin/13104_2016_1945_MOESM4_ESM.xlsx

S5: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782323/bin/13104_2016_1945_MOESM5_ESM.xlsx

S6: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4782323/bin/13104_2016_1945_MOESM6_ESM.xlsx

## REFERENCE LIST


1. Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, Harris J, Mallison GF, Martin SM, McDade JE, Shepard CC, Brachman PS: Legionnaires' disease: description of an epidemic of pneumonia. *N Engl J Med* 1977, 297:1189–1197.
2. Fry NK, Alexiou-Daniel S, Bangsborg JM, Bernander S, Castellani Pastoris M, Etienne J, Forsblom B, Gaia V, Helbig JH, Lindsay D, Christian Luck P, Pelaz C, Uldum SA, Harrison TG: A multicenter evaluation of genotypic methods for the epidemiologic typing of Legionella pneumophila serogroup 1: results of a pan-European study. *Clin Microbiol Infect* 1999, 5: 462–477.
3. Chiarini A, Bonura C, Ferraro D, Barbaro R, Calà C, Distefano S, Casuccio N, Belfiore S, Giammanco A: Genotyping of Legionella pneumophila serogroup 1 strains isolated in Northern Sicily, Italy. *New Microbiol* 2008, 31:217–28.
4. Doleans A, Aurell H, Reyrolle M, Lina G, Freney J, Vandenesch F, Etienne J, Jarraud S: Clinical and Environmental Distributions of Legionella Strains in France Are Different. *J Clin Microbiol* 2004, 42:458–460.
5. Den Boer JW, Bruin JP, Verhoef LPB, Van der Zwaluw K, Jansen R, Yzerman EPF: Genotypic comparison of clinical Legionella isolates and patient-related environmental isolates in The Netherlands, 2002-2006. *Clin Microbiol Infect* 2008, 14:459–466.
6. Harrison TG, Afshar B, Doshi N, Fry NK, Lee J V.: Distribution of Legionella pneumophila serogroups, monoclonal antibody subgroups and DNA sequence types in recent clinical and environmental isolates from England and Wales (2000-2008). *Eur J Clin Microbiol Infect Dis* 2009, 28:781–791.
7. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat RevGenet* 2008, 9:356–369.
8. Saeys Y, Inza I, Larra??aga P: A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007:2507–2517.
9. Guyon I, Elisseeff A: An introduction to variable and feature selection. *J Mach Learn Res* 2003, 3:1157–1182.
10. Wang HY, Zheng H, Azuaje F: Evaluation of computational classification methods for discriminating human heart failure etiology based on gene expression data. *2006 Comput Cardiol* 2006.
11. Guyon I, Weston J, Barnhill S, Vapnik V: Gene selection for cancer classification using support vector machines. *Mach Learn* 2002, 46:389–422.
12. Diaz-Uriarte R: GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics* 2007, 8:328.
13. Haury A-C, Gestraud P, Vert J-P: The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One* 2011, 6:e28210.

14. Diaz-Diaz N, Aguilar-Ruiz JS, Nepomuceno JA, Garcia J: Feature selection based on bootstrapping. In *Comput Intell Methods Appl 2005 ICSC Congr*; 2005.

15. Duangsoithong R, Windeatt T: Bootstrap feature selection for ensemble classifiers. In *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). Volume 6171 LNAI*; 2010:28–41.

16. Efron B, Tibshirani R: [Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy]: Rejoinder. *Stat Sci* 1986:77–77.

17. Hinkley D V: Bootstrap methods. *J R Stat Soc Ser B* 1988, 50:321–337.

18. John G, Kohavi R, Pfleger K: Irrelevant Features and the Subset Selection Problem. *ICML* 1994:121–129.

19. Kohavi R: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Int Jt Conf Artif Intell. Volume 14*; 1995:1137–1143.

20. Harrell FE: Model uncertainty, penalization, and parsimony. *ISCB Present UVa Web page* 1998.

21. Austin PC, Tu J V: Bootstrap Methods for Developing Predictive Models. *Am Stat* 2004: 131–137.

22. Roberts S, Martin MA: Bootstrap-after-bootstrap model averaging for reducing model uncertainty in model selection for air pollution mortality studies. *Environ Health Perspect* 2010, 118:131–136.

23. Breiman L, Friedman JH, Olshen RA, Stone CJ: *Classification and Regression Trees. Volume 19*; 1984.

24. Breiman LEO: Random Forests. *Mach Learn* 2001, 45:5–32.

25. Cortes C, Vapnik V: Support-vector networks. *Mach Learn* 1995, 20:273–297.

26. Tibshirani R: Regression shrinkage and selection via the lasso: A retrospective. *J R Stat Soc Ser B Stat Methodol* 2011, 73:273–282.

27. Euser SM, Nagelkerke NJ, Schuren F, Jansen R, Den Boer JW: Genome Analysis of Legionella pneumophila Strains Using a Mixed-Genome Microarray. *PLoS One* 2012, 7.

28. R Development Core Team R: R: A Language and Environment for Statistical Computing. *R Found Stat Comput* 2011:409.

29. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF: Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001, 54:774–781.

# Chapter 6

## Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints

Tjeerd van der Ploeg; Peter C. Austin, Ewout W. Steyerberg

BMC Med Res Methodol. 2014 Dec 22;14:137. doi: 10.1186/1471-2288-14-137.

## ABSTRACT

### Background

Modern modelling techniques may potentially provide more accurate predictions of binary outcomes than classical techniques. We aimed to study the predictive performance of different modelling techniques in relation to the effective sample size ("data hungriness").

### Methods

We performed simulation studies based on three clinical cohorts: 1282 patients with head and neck cancer (with 46.9% 5 year survival), 1731 patients with traumatic brain injury (22.3% 6-month mortality) and 3181 patients with minor head injury (7.6% with CT scan abnormalities). We compared three relatively modern modelling techniques: support vector machines (SVM), neural nets (NN), and random forests (RF) and two classical techniques: logistic regression (LR) and classification and regression trees (CART). We created three large artificial databases with 20 fold, 10 fold and 6 fold replication of subjects, where we generated dichotomous outcomes according to different underlying models. We applied each modelling technique to increasingly larger development parts (100 repetitions). The area under the ROC-curve (AUC) indicated the performance of each model in the development part and in an independent validation part. Data hungriness was defined by plateauing of AUC and small optimism (difference between the mean apparent AUC and the mean validated AUC <0.01).

### Results

We found that a stable AUC was reached by LR at approximately 20 to 50 events per variable, followed by CART, SVM, NN and RF models. Optimism decreased with increasing sample sizes and the same ranking of techniques. The RF, SVM and NN models showed instability and a high optimism even with >200 events per variable.

### Conclusions

Modern modelling techniques such as SVM, NN and RF may need over 10 times as many events per variable to achieve a stable AUC and a small optimism than classical modelling techniques such as LR. This implies that such modern techniques should only be used in medical prediction problems if very large data sets are available.

## 6.1 INTRODUCTION

Prediction of binary outcomes is important in medical research. The interest in the development, validation, and clinical application of clinical prediction models is increasing [1]. Most prediction models are based on logistic regression analysis (LR), but other, more modern techniques, may also be used. Support vector machines (SVM), neural nets (NN) and random forest (RF) have received increasing attention in medical research [2] [3] [4] [5] [6], since these hold the promise of better capturing non-linearities and interactions in medical data. The increased flexibility of modern techniques implies that larger sample sizes may be required for reliable estimation. Little is known, however, about the sample size that is needed to generate a prediction model with a modern modelling technique that outperforms more traditional, regression-based modelling techniques in medical data.

Usually, only a relatively limited number of subjects is available for developing prediction models. In 1995, a comparative study on the performance of various prediction models for medical outcomes concluded that the ultimate limitation seemed due to the availability of the information in data. This study used the term "data barrier" [7].

Some researchers aimed to develop a "power law" that can be used to determine the relation between sample size and the discriminatory ability of prediction models in terms of accuracy [8] [9] [10]. These studies clarified how a larger sample size leads to a better accuracy. The studies revealed that a satisfactory level of accuracy (the accuracy at infinite sample size +/- 0.01) can be achieved by sample sizes varying from 300 to 16,000 records, depending on the modelling technique and the data structure. The relation between sample size and accuracy was reflected in learning curves. Similarly, the number of events per variable (EPV) has been studied in relation to model performance [11] [12] [13] [14] [15].

In the current study, we aimed to define learning curves to reflect the performance of a model in terms of discriminatory ability, which is a key aspect of the performance of prediction models in medicine [16]. We assumed that the discriminatory ability of a model is a monotonically increasing function of the sample size, converging to a maximum at the infinite sample size. We hypothesized that modern, more flexible techniques are more "data hungry" [17] than more conventional modelling techniques, such as regression analysis. The concept of data hungriness refers to the sample size needed for a modelling technique to generate a prediction model with a good predictive accuracy. For fair comparison, we generated reference models with each of the modelling techniques considered in our simulation study.

## 6.2 METHODS

### Patients

We performed a simulation study, based on three patient cohorts.
The first cohort consisted of patients with head and neck cancer who were followed during 15 years for survival ("HNSCC cohort") [18]. The cohort contained 7 predictor variables (2 dichotomous, 4 categorical and 1 continuous) and a dichotomous (0/1) outcome with an incidence of 601/1282 (46.9%).
The second cohort consisted of patients with traumatic brain injury ("TBI cohort") [19]. The cohort contained 10 predictor variables (4 dichotomous, 1 categorical and 4 continuous) and a dichotomous outcome with an incidence of 386/1731 (22.3%).
The third cohort consisted of patients suspected of head injury who underwent a CT-scan ("CHIP cohort") [6]. This cohort contained 12 predictor variables (9 dichotomous, 1 categorical and 2 continuous) and a dichotomous (0/1) outcome with an incidence of 243/3181 (7.6%).
We generated artificial cohorts by replicating the HNSCC cohort 20 times, the TBI cohort 10 times and the CHIP cohort 6 times. This resulted in an artificial cohort consisting of 25,640 subjects ("HNSCC artificial cohort"), an artificial cohort consisting of 17,310 subjects ("TBI artificial cohort") and an artificial cohort consisting of 19,086 subjects ("CHIP artificial cohort").

**Table 1** Cohort characteristics

| | **Cohort** | | |
|---|---|---|---|
| | HNSCC | TBI | CHIP |
| Outcome | 5 year survival | 6-months mortality | intracranial findings |
| Type | dichotomous | dichotomous | dichotomous |
| Event/Total | 601/1282 (46.9%) | 386/1731 (22.3%) | 243/3181 (7.6%) |
| Predictors | 2 dichotomous | 4 dichotomous | 9 dichotomous |
| | 4 categorial | 1 categorial | 1 categorial |
| | 1 continuous | 4 continuous | 2 continuous |

### Reference models

In the current study, we evaluated the following modelling techniques, using default settings as far as possible:

- Logistic regression (LR)
- Classification and regression trees (CART)
- Support vector machines (SVM)
- Neural nets (NN)
- Random forest (RF)

For a description of these modelling techniques, based on the work of various authors [12] [15] [20] [21], we refer to Appendix 1.
As reference points for this evaluation, we first applied each modelling technique to each entire artificial cohort in order to generate an LR model, a CART model, an SVM model, an NN model and an RF model. These models were fitted with optimization according to default settings. Next, we generated probabilities of the outcome for each of these reference models. With these probabilities, we generated a new 0/1 outcome by comparing the generated probabilities of each reference model with a random number from a uniform (0,1) distribution. Using this new 0/1 outcome, we evaluated the five modelling techniques. The R-code for the construction of the reference models is in Appendix 2.

### Development and validation

For each of the five modelling techniques, we randomly divided the artificial cohort into a development set and a validation set for performance assessment. Each set consisted of 50% of the subjects of the artificial cohort.

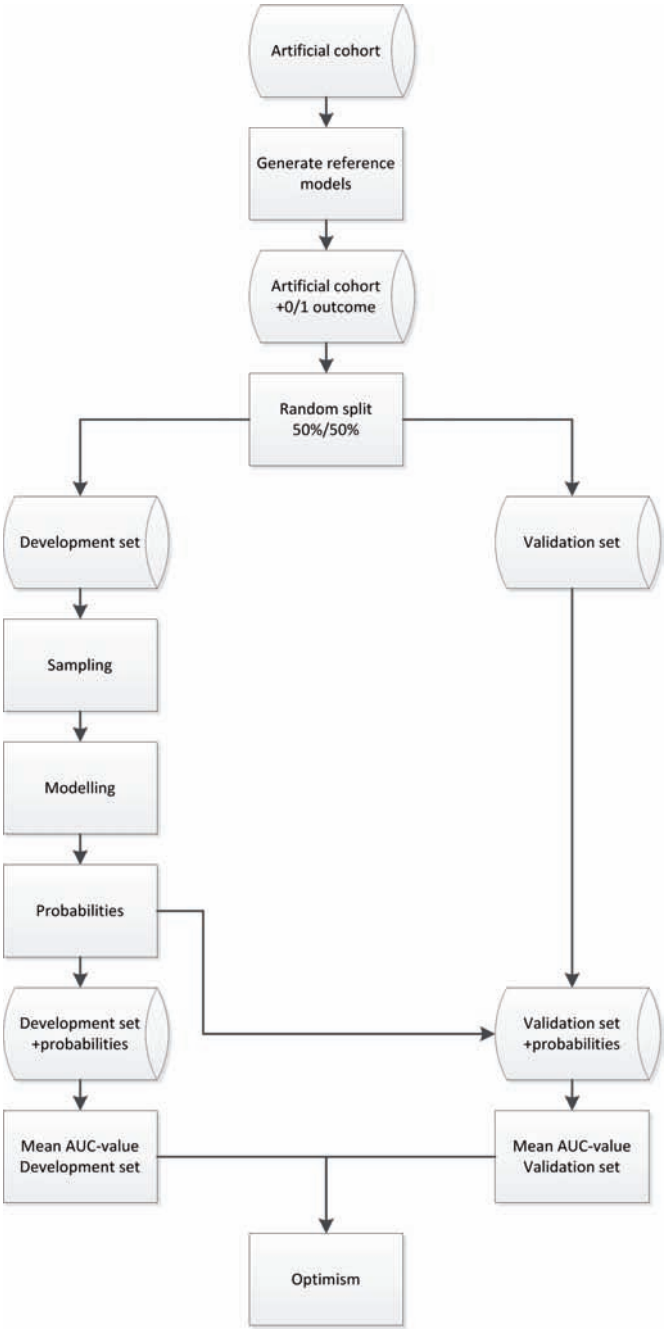
### Simulation design and analysis

We applied the following steps to each of the three artificial cohorts:

1. Development sets were samples of increasing sizes (varying from 200 to the maximum size of the development set with increment 1000), drawn at random from the non-validation part of the artificial cohort.
2. For each of the five modelling techniques we generated a model for each sample, taking the 0/1 outcome of a specific reference model as outcome. We evaluated the predictions on each sample.
3. For each sample, the predictions of the model were evaluated on the validation set, taking the 0/1 outcome of the same reference model as outcome.

We repeated these steps 100 times for each sample size to achieve sufficient stability. We considered each of the five reference models in turn for a fair comparison of each of the modelling techniques. Evaluation of predictive performance focussed on the discriminatory ability according to the area under the Receiver Operating Characteristic curve (AUC). The AUC was determined using the development set (apparent AUC) and the validation set (validated AUC). We calculated optimism as mean apparent AUC minus mean validated AUC.
We defined the maximally attainable AUC (AUCmax) as the validated AUC-value of a model based on the entire development set (50% of the artificial cohort).

**Figure 1** Flow chart simulation design



A flowchart of the simulation design is presented in Figure 1. For the analysis we used R software (version 2.14) [22]. For the R-code of the simulation design we refer to Appendix 2 [23].

### Learning curves

For each modelling technique, we generated learning curves to visualize the relation between the AUC-values and optimism of the generated models with respect to the number of events per variable.

### Data hungriness

The data hungriness of a modelling technique was defined as the minimum number of events per variable at which the optimism of the generated model was <0.01. This limit was admittedly arbitrary, but in line with previous research [24].

### Sensitivity analysis

We performed a sensitivity analysis to determine the influence of the endpoint incidence in the CHIP artificial cohort (7.6%). We hereto selectively oversampled subjects with the outcome of interest in order to generate an artificial cohort with an endpoint incidence of 50% ("CHIP5050 cohort").

## 6.3 RESULTS

### HNSCC cohort

The best performance in terms of mean validated AUC-values was achieved when the full development set was used (n=12,820, number of events=6013, event rate 46.9%) and by the models generated with the same modelling technique as the reference model, except when the reference model was generated with NN, in which case the RF model had the best performance (AUC 0.810, Table 2).
The level that could be reached (AUCmax) depended foremost on the reference model used to generate the 0/1 outcomes. All models performed best when the reference model RF was used. For all reference models, except the CART reference model, the CART model performed worst (Table 2).

The data hungriness of the various modelling techniques is reflected by the first part of the learning curves with <100 events per variable (Figure 2). As expected, all models converged monotonically to AUCmax. For each of the reference models, the LR model showed the most rapid increase to a stable mean validated AUC-value, while the RF

**Table 2** AUCmax per reference model, HNSCC cohort

| | Reference model | | | | |
|---|---|---|---|---|---|
| | **LR** | **CART** | **SVM** | **NN** | **RF** |
| **LR** | **0.797** | 0.745 | 0.803 | 0.802 | 0.880 |
| **CART** | 0.730 | **0.748** | 0.749 | 0.728 | 0.822 |
| **SVM** | 0.787 | 0.740 | **0.814** | 0.802 | 0.898 |
| **NN** | 0.785 | 0.745 | 0.800 | **0.804** | 0.869 |
| **RF** | 0.784 | 0.747 | 0.810 | 0.810 | **0.929** |

**Bold numbers** are for model performance when the underlying model was specified according to the modelling technique considered.

model needed the largest number of events per variable to reach a stable mean validated AUC-value (Figure 2).

We calculated the relative performance of a model by setting the performance of the model resulting from the modelling technique that generated the reference model at 100%. Figure 3 shows the relative performance of the models for each reference model.

**Figure 2** Validated AUC-values vs. events per variable, HNSCC cohort

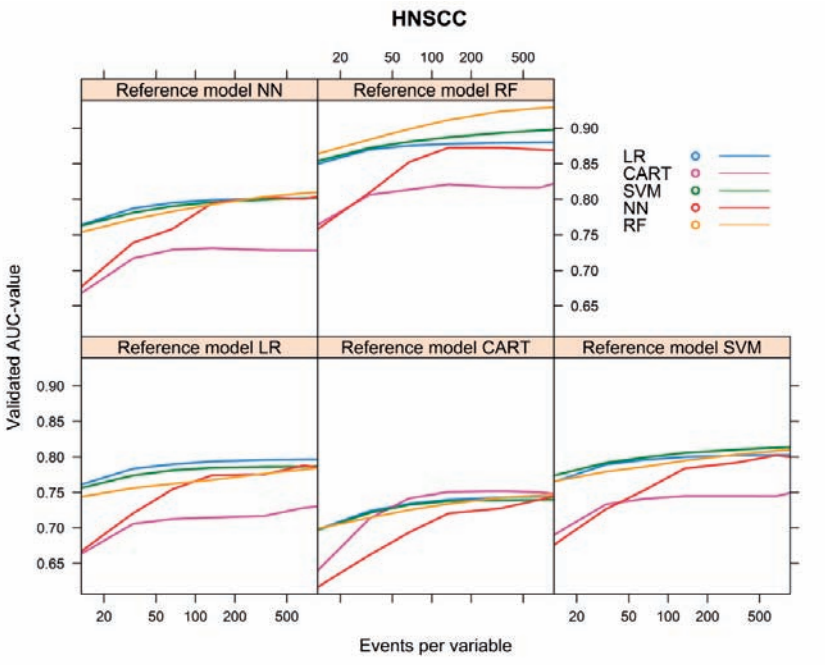
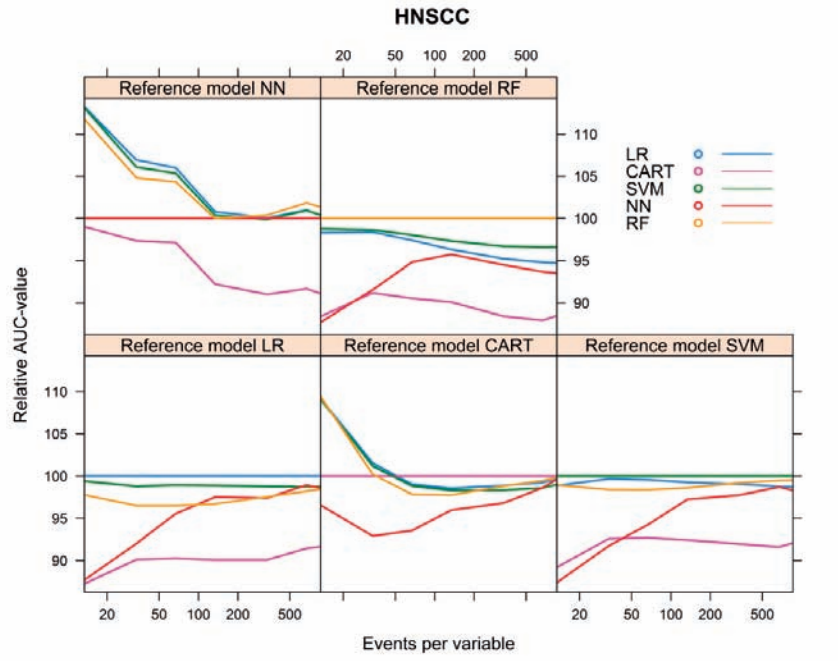


**Figure 3** Relative validated AUC-values vs. events per variable, HNSCC cohort




For all reference models, the optimism of the models decreased with an increasing number of events per variable. For all reference models, except when the reference model was CART, the modelling technique LR needed the smallest number of events per variable to reach an optimism <0.01 (55 to 127 events per variable).

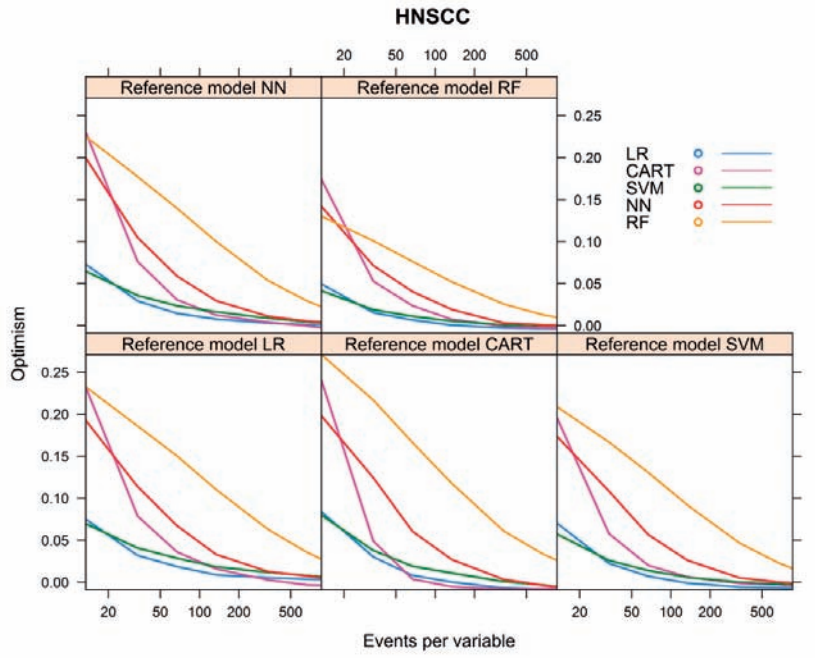
When CART was the reference model, the modelling technique CART needed the smallest number of events per variable to reach an optimism <0.01 (62 events per variable). The modelling techniques NN and RF and, to a lesser extent, SVM needed the most events per variable to generate models with an optimism <0.01.

The modelling technique RF needed 850 events per variable when the reference model RF was used, but for the other reference models the optimism of the RF model remained >=0.01, despite the large number of events per variable (Figure 4).

## TBI cohort

For the TBI artificial cohort, with a development set consisting of 8655 subjects and 1930 events (event rate 22.3%), the CART models performed poorly, irrespective of the reference model (Table 3). The models generated with the same modelling technique as the reference model showed the best performance, except when the reference model was generated with CART, in which case the LR model had the best performance (AUC

**Figure 4** Optimism vs. events per variable, HNSCC cohort




0.712, Table 3). All models, except the CART model, showed the lowest AUC when the reference model CART was used (Table 3).

The NN model needed the largest number of events per variable to reach AUCmax. For each of the reference models, the LR model showed the most rapid increase to a stable AUC (Figure 5).

Again, we calculated the relative performance of a model by setting the performance of the model resulting from the modelling technique that generated the reference model at 100%. Figure 6 shows the relative performance of the models for each reference model.

For all models, optimism decreased with an increasing number of events per variable. The LR model needed 18-23 events per variable to reach an optimism <0.01, whereas the optimism of the RF model remained high, except for the reference model RF, in which case optimism was <0.01 at 163 events per variable (Figure 7).

**Table 3** AUCmax per reference model, TBI cohort

| | Reference model | | | | |
|---|---|---|---|---|---|
| | **LR** | **CART** | **SVM** | **NN** | **RF** |
| **LR** | **0.806** | 0.712 | 0.743 | 0.792 | 0.817 |
| **CART** | 0.710 | **0.702** | 0.676 | 0.652 | 0.684 |
| **SVM** | 0.754 | 0.677 | **0.765** | 0.765 | 0.838 |
| **NN** | 0.800 | 0.701 | 0.746 | **0.802** | 0.828 |
| **RF** | 0.744 | 0.685 | 0.750 | 0.776 | **0.988** |

**Bold numbers** are for model performance when the underlying model was specified according to the modelling technique considered.

**Figure 5** Validated AUC-values vs. number of events per variable, TBI cohort

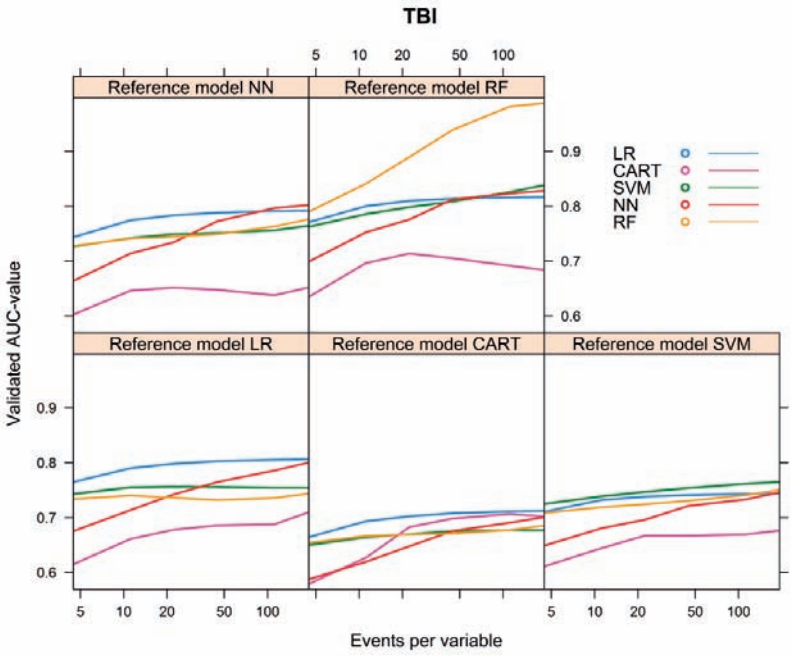


**Figure 6** Relative validated AUC-values vs. events per variable, TBI cohort

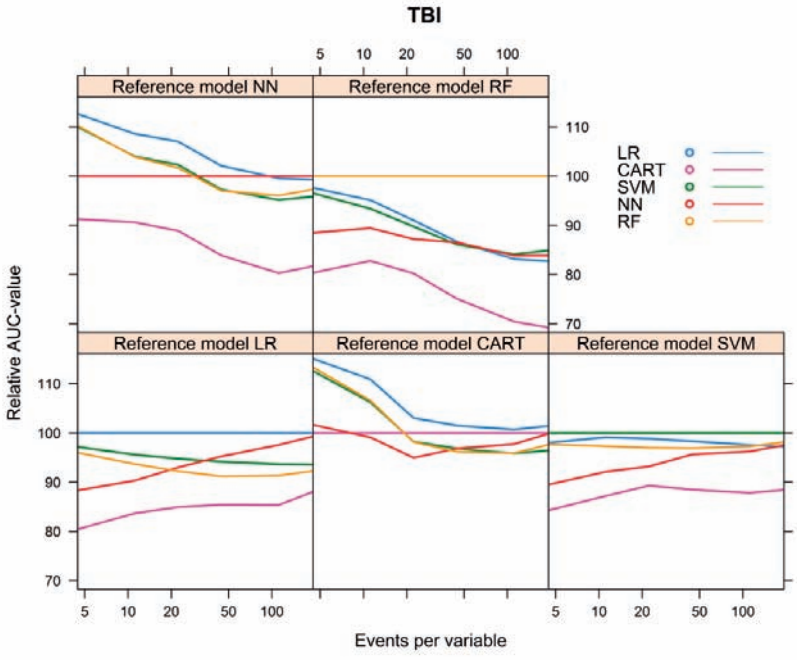
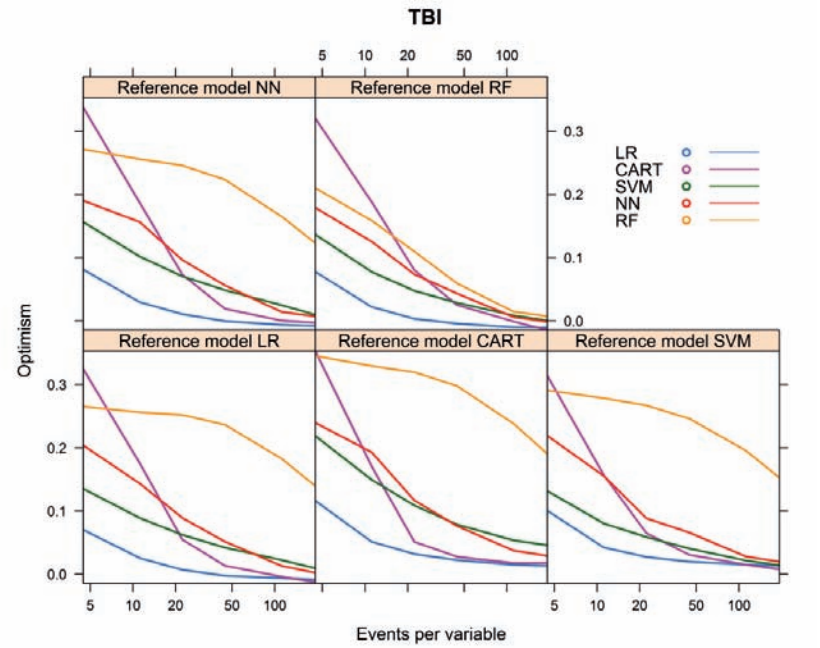



**Figure 7** Optimism vs. events per variable, TBI cohort



## CHIP cohort

For the CHIP artificial cohort, with a development set consisting of 9543 subjects and 729 events (event rate 7.64%), the findings were largely similar to the results of the HNSCC cohort. The best performance was achieved by the same modelling technique that generated the reference model (Table 4). The modelling technique CART generated models with a poor performance, irrespective of the reference models. The modelling technique SVM also generated models with a poor performance, irrespective of the reference models, except when the RF model was used as reference model (AUC 0.871, Table 4). All models performed poorly when the reference models CART and SVM were used. All models, except the CART model, performed well when the reference model RF was used (AUC>0.8, Table 4).

**Table 4** AUCmax per reference model, CHIP cohort

| | Reference model | | | | |
|---|---|---|---|---|---|
| | **LR** | **CART** | **SVM** | **NN** | **RF** |
| **LR** | **0.786** | 0.572 | 0.607 | 0.782 | 0.903 |
| **CART** | 0.562 | **0.578** | 0.580 | 0.500 | 0.666 |
| **SVM** | 0.584 | 0.560 | **0.615** | 0.616 | 0.871 |
| **NN** | 0.758 | 0.564 | 0.589 | **0.791** | 0.856 |
| **RF** | 0.728 | 0.579 | 0.594 | 0.755 | **0.916** |

**Bold numbers** are for model performance when the underlying model was specified according to the modelling technique considered.

Considering the learning curves (Figure 8), the CART models performed poorly. For each of the reference models, the LR model showed a rapid increase to a stable mean validated AUC-value, in contrast to the NN model which needed far more events to reach a stable mean validated AUC-value. The CART model showed a decreasing mean validated AUC-value despite increasing number of events, except when the reference model CART was used (Figure 8).

Figure 9 shows the relative performance of the models for each reference model. For the reference models LR, SVM and NN, the modelling technique LR required 14 to 28 events per variable to reach an optimism <0.01 and CART required 11 to 17 events per variable. Despite an increasing number of events per variable, the modelling techniques SVM, NN and RF generated models with optimism >0.01 for all reference models. For the reference models CART and RF, none of the modelling techniques was able to generate a model with optimism <0.01 (Figure 10).

**Figure 8** Validated AUC-values vs. number of events per variable, CHIP cohort

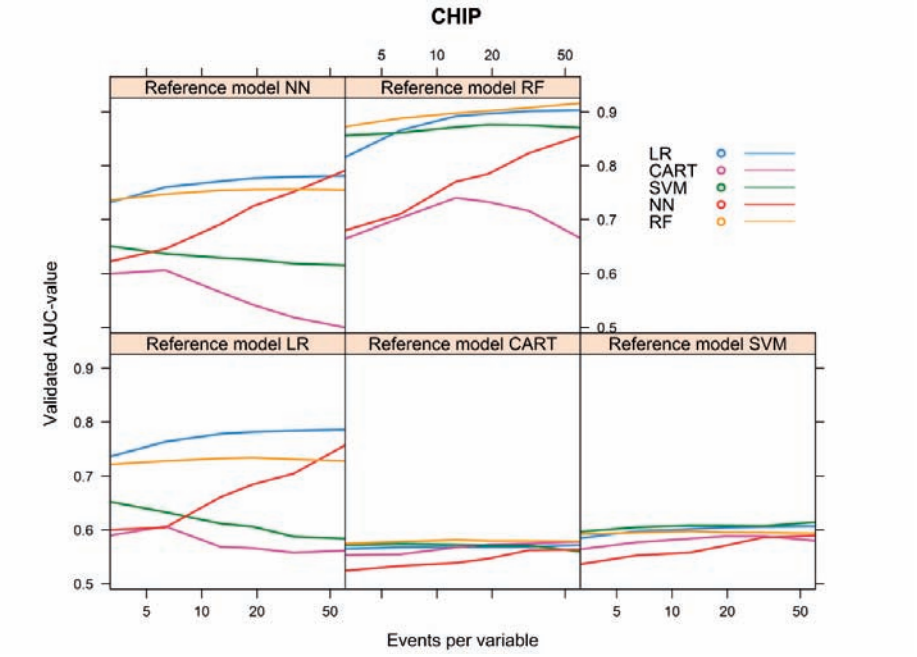



**Figure 9** Relative validated AUC-values vs. events per variable, CHIP cohort

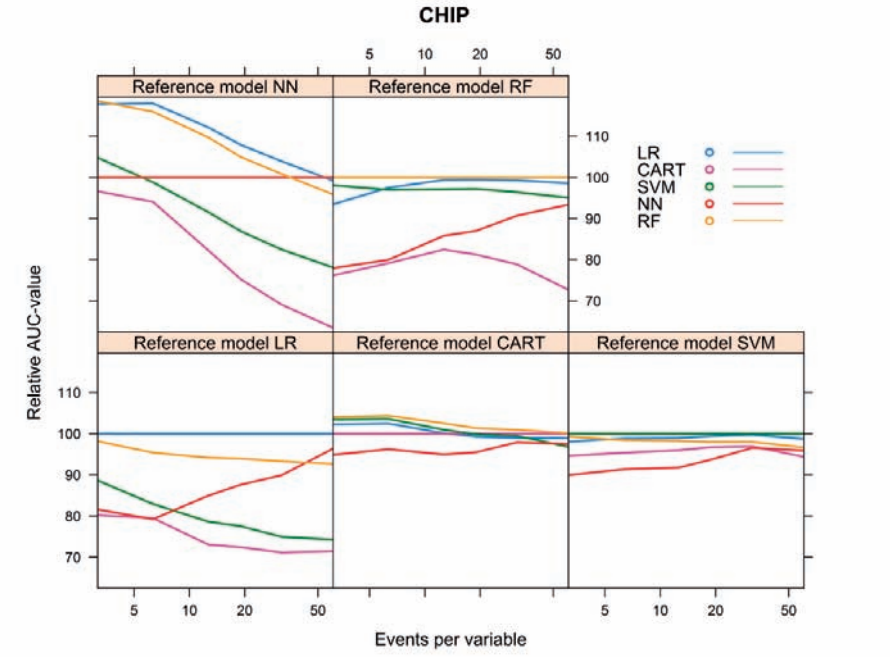
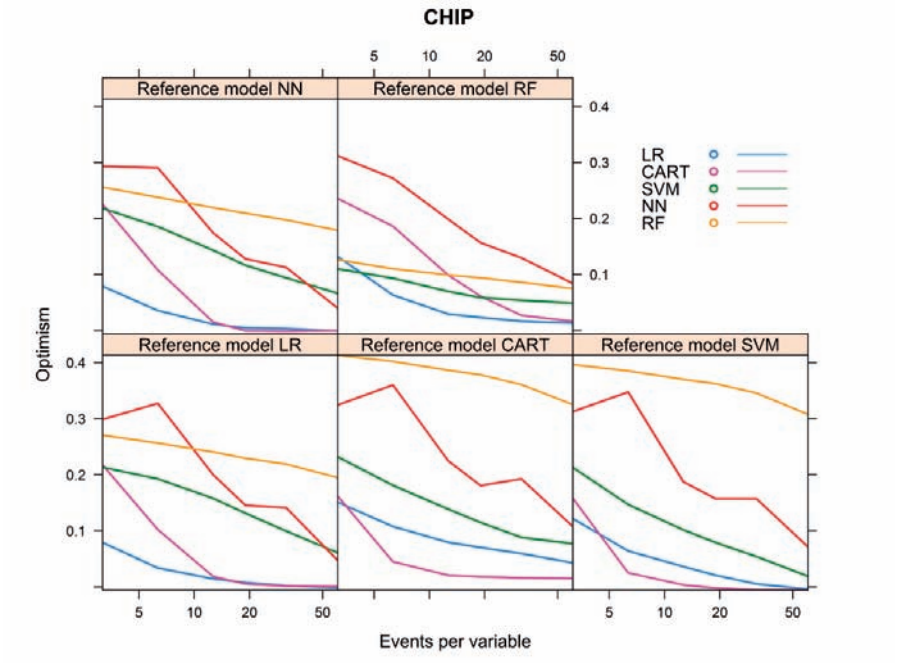


**Figure 10** Optimism vs. events per variable, CHIP cohort




### Sensitivity analysis CHIP cohort

When we increased the event rate in the CHIP cohort from 7.6% to 50% ("CHIP5050 cohort"), the behaviour of the learning curves became largely similar to the behaviour of the curves generated for the HNSCC cohort (Appendix 3, Figure 11 to 13).

## 6.4 DISCUSSION

Modern modelling techniques, such as SVM, NN and RF, needed far more events per variable to achieve a stable validated AUC and an optimism <0.01 than the more conventional modelling techniques, such as LR and CART. The CART models had a stable performance, but at a fairly poor level. Specifically, a larger number of events did not lead to better validated performance in the cohort with a 7.6% event rate. The LR models had low optimism when the number of events per variable was at least 20 to 50. A remarkable finding was that the optimism of the RF models remained high for the three cohorts, even at a large number (over 200) of events per variable. This indicates that these RF models were far from robust. Of note, the validated performance of RF models was similar to that of LR models. This implies that especially RF models need

careful validation to assess predictive performance, since apparent performance may be highly optimistic.

Since LR modelling is far less data hungry than alternative modelling techniques, this technique may especially be useful in relatively small data sets. With very small data sets, any modelling technique will lead to poorly performing models. Our results confirm the generally accepted rule that reasonable predictive modelling requires at least 10 events per variable, even with a robust technique such as LR [11] [12] [15]. We note that larger numbers of events per variable are desirable to achieve better stability and higher expected performance.

The modelling techniques SVM and NN needed far more events per variable to generate models with a stable mean validated AUC-value and an optimism converging towards zero. For models generated with the modelling technique RF, the optimism did not even converge towards zero at the largest number of events per variable that we evaluated. Obviously, models generated by the same modelling technique as the reference model generally performed best, reflecting a "home advantage" over models generated by a different modelling technique than the reference model. The performance of models according to different reference models was provided for a fair assessment of the performance of the approaches considered.

While RF and LR models consistently performed well, CART consistently performed poorly. The poor performance of CART modelling may be explained by the fact that continuous variables need to be categorized, with optimal cut-offs determined from all possible cut-off points, and that possibly unnecessary higher-order interactions are assumed between all predictor variables. RF modelling is an obvious improvement over CART modelling [24]. It is hence remarkable that CART is still advocated as the preferred modelling technique for prediction in some disease areas, such as trauma [25]. A researcher must always carefully consider which modelling technique is appropriate in a specific situation. Using, for instance, a random forest technique just because the number of subjects is over 10,000 is too simplistic.

The aim of our study was to investigate the data hungriness of the various modelling techniques and the aim was not to find the best modelling technique in AUC terms. To our knowledge, the data hungriness of various modelling techniques has not been assessed before for medical prediction problems. However, a few studies addressed this topic in the context of progressive sampling for the development of a power law to guide the required sample size for prediction modelling. For example, arithmetic sampling was applied with sample sizes of 100, 200, 300, 400 etc. to 11 of the UCI repository databases to obtain insight into the performance of a naive Bayes classifier [8]. This study led to required sample sizes from 300 to 2180 to be within 2% from the accuracy of a model built from the entire database. Other researchers modelled 3 of the larger databases from the UCI repository using different progressive sampling

techniques [9]. Using the C4.5 modelling technique, which we consider a CART variant, sample sizes of 2000 for the LED database, 8000 for the CENSUS database and 12000 for the WAVEFORM database were required for a model being no more than 1% less accurate than a model based on all the available data.

Another study compared the performance of 6 data mining tools at various sample sizes for 2 test databases (test database I with 50,000 records and test data base II with 1,500,000 records), using accuracy as the performance measure. For test database I, for all tools, a stable level of accuracy was reached at 16,000 records, and for test database II, for all tools, a stable level was reached at 8000 records [10]. The results of our study are in line with these studies. Although we used mean validated AUC-values instead of accuracy to measure the performance of the models, we also found that the more complex modelling techniques required large numbers of events per variable to generate models with optimism <0.01.

A number of limitations need to be considered. Firstly, we used three cohorts with dichotomous outcomes, in which non-linearity was not a major issue. While this may be common in medical research, it limited the ability for some modern modelling techniques to outperform traditional logistic regression modelling. If important non-linearity is truly present in a data set, techniques that capture such non-linear patterns well are obviously attractive. Various approaches can be considered to address non-linearity within the regression framework, including restricted cubic splines and fractional polynomials [15] [26]. Secondly, we used default settings for the modelling techniques [8]. Further research might investigate our evaluated models, but also other modelling techniques such as LASSO, using other cohorts, and also using other settings for the modelling (such as pruning options, priors, and number of subjects in the end nodes).

Thirdly, there was a considerable difference in incidence between the three cohorts (47%, 22% and 8%). To assess the effect of this difference in incidence on the data hungriness, we performed a sensitivity analysis. Further research should evaluate the relation between the incidence of the outcome and the data hungriness patterns of various modelling techniques.

## Conclusions

Modern modelling techniques such as SVM, NN and RF need far more events per variable to achieve a stable AUC-value than classical modelling techniques such as LR and CART. If very large data sets are available, modern techniques such as RF may potentially achieve an AUC-value that exceeds the AUC-values of modelling techniques such as LR. The improvement over simple LR models may, however, be minor, as was shown in the two empirical examples in this study. This implies that modern modelling

techniques should only be considered in medical prediction problems if very large data sets with many events are available.

## 6.5 ABBREVIATIONS

CT: Computed tomography; SVM: Support vector machines; NN: Neural nets; RF: Random forest; CART: Classification and regression trees; LR: Logistic regression; ROC: Receiver operating curve: AUC: Area under the curve; EPV: Events per variable; HNSCC: Head and neck squamous cell carcinoma; TBI: Traumatic brain injury; CHIP: CT in head injury patients; UCI: University of California, Irvine; LASSO: Least absolute shrinkage and selection operator.

## 6.6 ACKNOWLEDGEMENTS

The authors thank Mr. D. Nieboer for methodological and statistical advice and Mrs. L. van Hulst for editorial support.

## 6.7 APPENDIX 1

This appendix describes the evaluated modelling techniques in detail, based on the work of several authors [12] [15] [20] [21].

*Logistic regression (LR)*

Logistic regression is a type of regression analysis used for predicting the outcome of a binary dependent variable (a variable which can take only two possible outcomes, e.g. "yes" vs. "no" or "success" vs. "failure") based on one or more predictor variables. Logistic regression attempts to model the probability of a "yes/success" outcome using a linear function of the predictors. Specifically, the log-odds of success (the logit of the probability) is fit to the predictors using linear regression. Logistic regression is one type of discrete choice model, which in general predict categorical dependent variables, either binary or multi-way.

Like other forms of regression analysis, logistic regression makes use of one or more predictor variables that may be either continuous or categorical. Also, like other linear regression models, the expected value (average value) of the response variable is fit to the predictors, the expected value of a Bernoulli distribution is simply the probability of success. Unlike ordinary linear regression, however, logistic regression is used for predicting binary outcomes (Bernoulli trials) rather than continuous outcomes, and models a transformation of the expected value as a linear function of the predictors, rather than the expected value itself.

*Classification and regression trees (CART)*

Classification and regression trees is a tree-based classification and prediction modelling technique which uses recursive partitioning to split the training records into segments with similar output variable values. The modelling starts by examining the input variables to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two further subgroups and so on, until the stopping criterion is met. The parameter of RPART is the cp-parameter (cost complexity factor). A cp-value of 0.001 for example regulates that a split must decrease the overall lack of fit by a factor of 0.001.

*Support vector machine (SVM)*

A Support Vector Machine performs classification tasks by constructing hyperplanes with a margin in a multidimensional space that separates cases from different classes. SVM can efficiently perform a non-linear classification or regression task using different kernels (radial, linear and polynomial). The tuning parameters for SVM are the C-parameter (cost), which regulates the margin width, and the gamma-parameter for the

kernel calculation. SVM claims to be a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM may particularly be suited to analyse data with large numbers of predictor variables.

*Neural nets (NN)*

A neural network (NN), sometimes called a multilayer perceptron, works by simulating a large number of interconnected simple processing units, which are arranged in layers. There are three parts in a neural network: an input layer, with units representing the predictor variables, one or more hidden layers and an output layer, with a unit representing the outcome variable. The units are connected with varying connection strengths or weights. Input data are presented to the input layer and values are propagated from there to the next layer. Then, a prediction is delivered from the output layer. The network learns by examining individual records, generating a prediction for each record and making adjustments to the weights whenever it makes an incorrect prediction. This process is repeated many times, and the network continues to improve its predictions until one or more of the stopping criteria have been met. Initially, all weights are random, and the predictions that come out of the net are nonsensical. The network learns through training. Records for which the output is known are repeatedly presented to the network, and the predictions it gives are compared to the known outcomes. As training progresses, the network becomes increasingly accurate in replicating the known outcomes. Once trained, the network can be applied to new patients for whom the outcome is unknown. The parameters of NN are the size-parameter (number of units in the layer) and decay-parameter.

*Random forest (RF)*

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark.

Each tree is constructed using the following algorithm:

1. Let the number of training cases be N, and the number of variables in the classifier be M.
2. We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M.
3. Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.

4. For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set.
5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

## 6.8 APPENDIX 2

This appendix describes the R-code that was used for the simulation design on the HNSCC artificial cohort with a LR model as reference.

**# Open libraries**

```
library(foreign)
library(KsPlot)
```

**# Cohort creation**

```
HNSCC=read.spss("HNSCC20x.sav",use.value.labels=FALSE,to.data.frame=TRUE)
Gender=as.factor(HNSCC$Gender)
Tumor_location=as.factor(HNSCC$Tumor_location)
T_class=as.factor(HNSCC$T_class)
N_class=as.factor(HNSCC$N_class)
M_class=as.factor(HNSCC$M_class)
Prior_malignancies=as.factor(HNSCC$Prior_malignancies)
Age_at_diagnosis=as.numeric(HNSCC$Age_at_diagnosis)
ACE27=as.factor(HNSCC$ACE27)
Dead_or_alive_at_60_months=as.numeric(HNSCC$Dead_or_alive_at_60_months)
HNSCC2<-data.frame(Gender,Tumor_location,T_class,N_class,  Prior_malignancies,Age_at_
    diagnosis,ACE27,Dead_or_alive_at_60_months)
```

**# Construction of a binary outcome with the LR model as reference model**

```
lrModel <- glm(as.factor(Dead_or_alive_at_60_months)~ ., data = HNSCC2, family = "binomial")
lrProbs <- predict(lrModel, HNSCC2, type = "response")
lrROC <- caTools::colAUC(lrProbs,HNSCC2$Dead_or_alive_at_60_months)
lrROC
set.seed(1)
runis = runif(25640,0,1)
```

```
lry = ifelse(runis < lrProbs,1,0)

BASE<-data.frame(lry,Gender,Tumor_location,T_class,N_class,        Prior_malignancies,Age_at_
    diagnosis,ACE27)


# Creation development set and validation set


Sample <- sample(1:nrow(BASE), nrow(BASE)/2)

devBASE<- BASE[Sample, ]

valBASE<- BASE[-Sample, ]


# Modelling with the modelling techniques LR, CART, SVM, NN and RF with increasing sample
    size


output <- matrix(NA, nrow = 700, ncol=12, byrow=TRUE, dimnames = list(c(1:700),c("Sample
    number per size", "Sample size", "lrROCtraining","lrROCtest","cartROCtraining","cartRO
    Ctest","svmROCtraining","svmROCtest","nnROCtraining","nnROCtest","rfROCtraining",
    "rfROCtest")))

k=1

for( j in c(200, 500, 1000, 2000, 5000, 10000,nrow(devBASE)))

{

for (i in 1:100)

{

sampledata=devBASE[sample(1:nrow(devBASE),j),]

lrModel <- glm(as.factor(lry)~ ., data = sampledata, family = "binomial")

lrProbs1 <- predict(lrModel, sampledata, type = "response")

lrProbs2 <- predict(lrModel, valBASE, type = "response")

lrROCtraining<- caTools::colAUC(lrProbs1,sampledata$lry)

lrROCtest <- caTools::colAUC(lrProbs2,valBASE$lry)

cartModel <- mvpart::rpart(as.factor(lry)~., data = sampledata)

cartProbs1 <- predict(cartModel, sampledata)

cartProbs2 <- predict(cartModel, valBASE)

cartROCtraining<- caTools::colAUC(cartProbs1[,2],sampledata$lry)

cartROCtest <- caTools::colAUC(cartProbs2[,2],valBASE$lry)
```

```
svmModel <- e1071::svm(lry ~ ., data = sampledata,kernel = "polynomial", degree = 3, prob-
        ability = T)

svmProbs1 <- predict(svmModel, sampledata, probability = T)

svmProbs2 <- predict(svmModel, valBASE, probability = T)

svmROCtraining<- caTools::colAUC(svmProbs1,sampledata$lry)

svmROCtest <- caTools::colAUC(svmProbs2,valBASE$lry)

nnModel <- nnet::nnet(as.factor(lry) ~ ., data = sampledata, size = 10)

nnProbs1 <- predict(nnModel, sampledata)

nnProbs2 <- predict(nnModel, valBASE)

nnROCtraining<- caTools::colAUC(nnProbs1,sampledata$lry)

nnROCtest <- caTools::colAUC(nnProbs2,valBASE$lry)

rfModel <- randomForest::randomForest(lry ~ ., data = sampledata)

rfProbs1 <- predict(rfModel, sampledata)

rfProbs2 <- predict(rfModel, valBASE)

rfROCtraining<- caTools::colAUC(rfProbs1, sampledata$lry)

rfROCtest <- caTools::colAUC(rfProbs2, valBASE$lry)

output[k,]<-c(i,j,lrROCtraining,lrROCtest,cartROCtraining,cartROCtest,svmROCtraining,svmROCt
        est,nnROCtraining,nnROCtest,rfROCtraining, rfROCtest)

print(k)

k=k+1

}

}
```

**# Performance results to output file**

```
output

write.csv(output, "HNSCC training and test x vs lr.csv")
```

## 6.9 APPENDIX 3

This appendix shows the figures resulting from the sensitivity analysis.

**Figure 11** Validated AUC-values vs. events per variable, CHIP5050 cohort

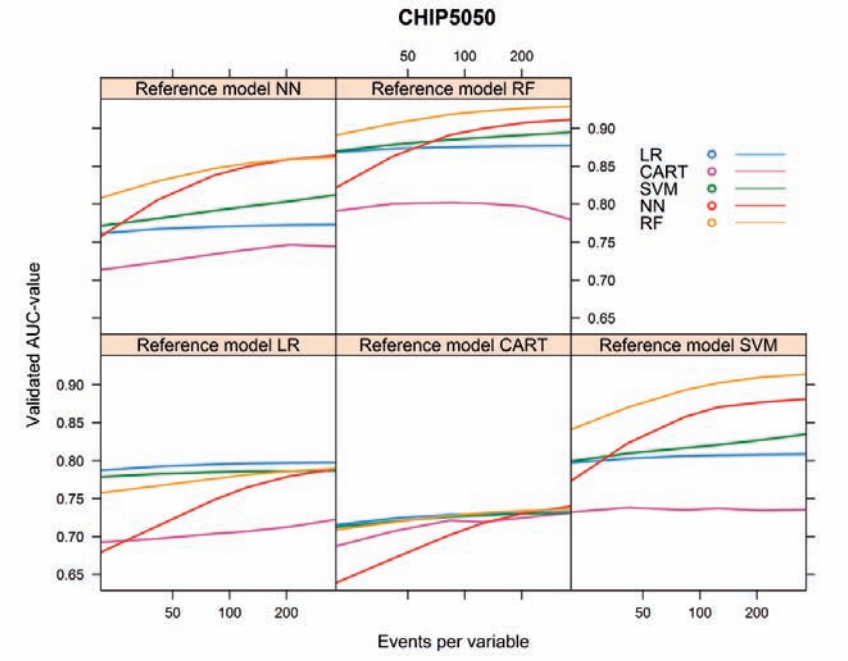


**Figure 12** Relative validated AUC-values vs. events per variable, CHIP5050 cohort

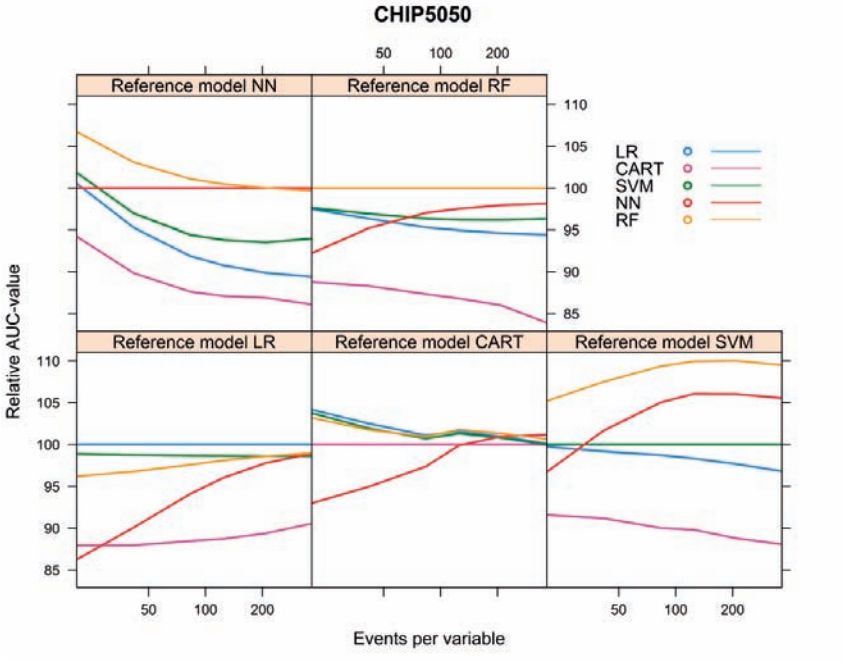
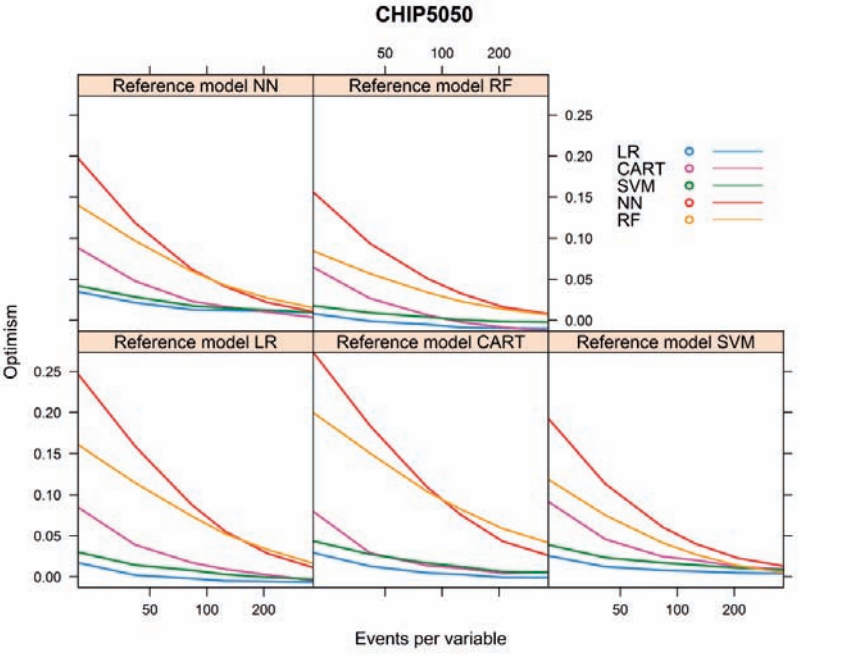



**Figure 13** Optimism vs. events per variable, CHIP5050 cohort



## REFERENCES

[1] E. W. Steyerberg, K. G. M. Moons, D. A. van der Windt, J. A. Hayden, P. Perel, S. Schroter, R. D. Riley, H. Hemingway, and D. G. Altman, "Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research," PLoS Med., vol. 10, 2013.

[2] R. F. Harrison and R. L. Kennedy, "Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation," Ann. Emerg. Med., vol. 46, pp. 431–439, 2005.

[3] Y.-J. L. O. L. Mangasarian and W. H. Wolberg, "Breast cancer survival and chemotherapy: a support vector machine analysis," in Discrete Mathematical Problems with Medical Applications: DIMACS Workshop Discrete Mathematical Problems with Medical Applications, December 8-10, 1999, DIMACS Center, 2000, vol. 55, p. 1.

[4] M. Ennis, G. Hinton, D. Naylor, M. Revow, and R. Tibshirani, "A comparison of statistical learning methods on the Gusto database.," Stat. Med., vol. 17, pp. 2501–2508, 1998.

[5] T. van der Ploeg, F. Datema, R. B. de Jong, and E. W. Steyerberg, "Prediction of Survival with Alternative Modelling Techniques Using Pseudo Values," PLoS One, vol. 9, no. 6, p. e100234, 2014.

[6] T. van der Ploeg, M. Smits, D. W. Dippel, M. Hunink, and E. W. Steyerberg, "Prediction of intracranial findings on CT-scans by alternative modelling techniques," BMC Medical Research Methodology, vol. 11. p. 143, 2011.

[7] H. P. Selker, J. L. Griffith, S. Patil, W. J. Long, and R. B. d'Agostino, "A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients.," J. Investig. Med. Off. Publ. Am. Fed. Clin. Res., vol. 43, no. 5, pp. 468–476, 1995.

[8] G. John, P. Langley, and H. John, "Static Versus Dynamic Sampling for Data Mining.," KDD, pp. 367–370, 1996.

[9] F. Provost, D. Jensen, and T. Oates, "Efficient Progressive Sampling," Proc. fifth ACM SIGKDD Int. Conf. Knowl. Discov. data Min., pp. 23–32, 1999.

[10] J. Morgan, R. Daugherty, A. Hilchie, and B. Carey, "Sample size and modelling accuracy of decision tree based data mining tools," Acad. Inf. Manag. Sci. J., vol. 6, no. 2, pp. 71–99, 2003.

[11] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis.," J. Clin. Epidemiol., vol. 49, pp. 1373–1379, 1996.

[12] F. E. Harrell, Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, 2001.

[13] E. W. Steyerberg, F. E. Harrell, G. J. J. M. Borsboom, M. J. C. Eijkemans, Y. Vergouwe, and J. D. F. Habbema, "Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis," J. Clin. Epidemiol., vol. 54, pp. 774–781, 2001.

[14] E. W. Steyerberg, M. J. C. Eijkemans, F. E. Harrell, and J. D. F. Habbema, "Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small data sets," Stat. Med., vol. 19, pp. 1059–1079, 2000.

[15] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, vol. 19. 2009, p. 500.

[16] I. Tzoulaki, G. Liberopoulos, and J. P. A. Ioannidis, "Assessment of claims of improved prediction beyond the Framingham risk score.," JAMA, vol. 302, pp. 2345–2352, 2009.

[17] D. B. Lee, "Requiem for large-scale models," ACM SIGSIM Simulation Digest, vol. 6. pp. 16–29, 1975.

[18] F. R. Datema, M. B. Ferrier, M. P. van der Schroeff, and R. J. Baatenburg de Jong, "Impact of comorbidity on short-term mortality and overall survival of head and neck cancer patients.," Head Neck, vol. 32, pp. 728–736, 2010.

[19] A. Marmarou, J. Lu, I. Butcher, G. S. McHugh, N. A. Mushkudiani, G. D. Murray, E. W. Steyerberg, and A. I. R. Maas, "IMPACT database of traumatic brain injury: design and description.," J. Neurotrauma, vol. 24, pp. 239–250, 2007.

[20] S. Tufféry, Data mining and statistics for decision making. John Wiley & Sons, 2011.

[21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, vol. 19. 1984, p. 368.

[22] R. R Development Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, vol. 1. p. 409, 2011.

[23] I. Kurahashi, "KsPlot: Check the power of seveal statistical models." 2011.

[24] P. C. Austin, D. S. Lee, E. W. Steyerberg, and J. V. Tu, "Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods?," Biometrical J., vol. 54, pp. 657–673, 2012.

[25] N. H. Young and P. J. D. Andrews, "Developing a Prognostic Model for Traumatic Brain Injury—A Missed Opportunity?," PLoS Med., vol. 5, p. 3, 2008.

[26] F. E. Harrell, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," Stat. Med., vol. 15, pp. 361–387, 1996.

# Chapter 7

## Modern modelling techniques had limited external validity in predicting mortality from traumatic brain injury

Tjeerd van der Ploeg; Daan Nieboer; Ewout W. Steyerberg

J Clin Epidemiol. 2016 Mar 14. pii: S0895-4356(16)30014-2. doi:10.1016/j.jclinepi.2016.03.002.

## ABSTRACT

### Background

Prediction of medical outcomes may potentially benefit from using modern statistical modelling techniques. We aimed to externally validate modelling strategies for prediction of 6-month mortality of patients suffering from traumatic brain injury (TBI) with predictor sets of increasing complexity.

### Methods

We analyzed individual patient data from fifteen different studies including 11,026 TBI patients. We consecutively considered a core set of predictors (age, motor score and pupillary reactivity), an extended set with CT scan characteristics, and a further extension with 2 laboratory measurements (glucose and hemoglobin). With each of these sets, we predicted 6-month mortality using default settings with five statistical modelling techniques: logistic regression (LR), classification and regression trees (CART), random forests (RF), support vector machines (SVM) and neural nets (NN). For external validation, a model developed on one of the fifteen data sets was applied to each of the fourteen remaining sets. This process was repeated fifteen times for a total of 630 validations. The area under the ROC-curve (AUC) was used to assess the discriminative ability of the models.

### Results

For the most complex predictor set, the LR models performed best (median validated AUC value 0.757), followed by RF and SVM models (median validated AUC value 0.735 and 0.732 respectively). With each predictor set, the CART models showed poor performance (median validated AUC value <0.7). The variability in performance across the studies was smallest for the RF and LR based models (IQR for validated AUC values from 0.07 to 0.10).

### Conclusions

In the area of predicting mortality from traumatic brain injury, non-linear and non-additive effects are not pronounced enough to make modern prediction methods beneficial.

## 7.1 INTRODUCTION

Prediction of binary outcomes has since long received much attention in medical research. Prediction is complicated by the specification of the model structure, such as the inclusion of main effects, potential non-linearities and statistical interactions [1] [2] [3]. While most prediction models for binary endpoints are still based on logistic regression analysis, there is increasing interest in other, more modern techniques, such as support vector machines, neural nets and random forests. These more modern methods hold the promise of better capturing non-linearities and interactions in medical data [4].

A decisive factor in choosing a modelling technique for prediction is the performance of the resulting model at external validation. Many studies compared modern modelling techniques with classical techniques, but mostly they only validated the resulting models internally [5] [6]. External validation was used in only a few comparisons of classification trees, neural networks and logistic regression [7] [8], and in a comparative study on stroke patients [9].

In this study, we aimed to compare the external validity of logistic regression and four more modern modelling techniques to predict 6-month mortality of patients suffering from traumatic brain injury (TBI). We chose this patient group because TBI is a heterogeneous disease, in which many mechanisms and pathways can lead to mortality and poor long term outcome [10] [11] [12] [13]. Moreover, tree-based models have specifically been suggested to be beneficial for prediction of outcome after TBI [4]. In patients with moderate or severe injuries, mortality 6-month after surgery exceeds 20% and lifelong disability occurs in half of the survivors [14]. Prediction of outcome in patients with TBI using prediction models has been studied since the 1970s [15] [16]. However, the preferred technique for prediction of outcome of TBI patients is still under debate, and preference for a technique varies between investigators [4]. Various statistical techniques have been used in this area, including logistic regression, recursive partitioning, Bayesian approaches and discriminant analysis [16]. Nowadays, a wide array of modern learning techniques is available, including random forests, support vector machines and neural networks [1] [17]. We investigated whether non-linear and non-additive effects in the area of predicting mortality from traumatic brain injury are pronounced enough such that these modern modelling techniques can outperform traditional modelling techniques such as logistic regression.

## 7.2 METHODS

### Patients

We analyzed individual patient data from the IMPACT database [14] [18] [19]. This database includes data of patients suffering from moderate or severe traumatic brain injury (TBI). The database comprises data from 11,026 patients included in fifteen different studies (Appendix 1, Table 1, Figure 1). Patients were enrolled in one of ten randomized clinical trials or in one of five registries between 1984 and 2006.

### Modelling techniques

We compared five statistical modelling techniques to predict 6-month mortality:
-Logistic regression (LR)
-Classification and regression trees (CART)
-Random forests (RF)
-Support vector machines (SVM)
-Neural nets (NN)

We here list the main characteristics of the evaluated modelling techniques, based on previous work of several authors [2] [3] [17] [20] [21]. We refer to Appendix 3 for the code of our analyses in R software [22].

*Logistic regression (LR)*

Logistic regression is a type of regression analysis that is often used in medical research to model the probability of a binary endpoint using a linear function of the predictors. Predictor variables may be either continuous or categorical. Logistic regression uses a logistic transformation to calculate the probability of a binary outcome. Regression coefficients were estimated by maximum likelihood using the *lrm* function in the *rms* library.

*Classification and regression trees (CART)*

CART is modelling technique that uses recursive partitioning to split the training records into segments with similar endpoint values. The modelling starts by examining the input variables to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two further subgroups and so on, until a stopping criterion is met. The commonly used parameter for CART is the cp-parameter (cost complexity factor). A cp-value of 0.001 for example regulates that a split must decrease the overall lack of fit by a factor of 0.001. The modelling was done using the *rpart* function in the *rpart* library.

### *Random forest (RF)*

Random forest is an ensemble classifier that consists of many decision trees. In case of classification, random forest outputs the class that is the mode among the classes from individual trees. In case of regression, random forest outputs the value that is the mean of the values output from individual trees. Each tree is constructed using a bootstrap sample from the original data. A tree is grown by recursively partitioning the bootstrap sample based on optimization of a *split rule*. In regression problems, the split rule is based on minimizing the mean squared error, whereas in classification problems, the Gini index is commonly used. At each split, a subset of candidate variables are tested for the split rule optimization, similar to CART modelling. For prediction, a new sample is pushed down the tree. This procedure is iterated over all trees in the ensemble. Key parameters are the number of trees and the number of candidate variables. The modelling was done using the *randomForest* function in the *randomForest* library.

### *Support vector machine (SVM)*

A SVM performs classification tasks by constructing hyperplanes with a margin in a multidimensional space that separates cases from different classes. A SVM can perform a non-linear classification or regression task using different kernels (radial, linear and polynomial). The tuning parameters for SVMs are the C-parameter (cost), which regulates the margin width, and the gamma-parameter for the kernel calculation. SVM claims to be a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM may particularly be suited to analyse data with large numbers of predictor variables. The modelling was done using the *svm* function in the *e1071* library.

### *Neural nets (NN)*

A NN, sometimes called a multilayer perceptron, simulates a large number of interconnected simple processing units, which are arranged in layers. There are three parts in a neural network: an input layer, with units representing the predictor variables, one or more hidden layers, and an output layer, with a unit representing the endpoint. The units are connected with varying connection strengths or weights. Input data are presented to the input layer and values are propagated from there to the next layer. Then, a prediction is delivered from the output layer. The network learns by examining individual records, generating a prediction for each record and making adjustments to the weights whenever it makes an incorrect prediction. The adjustments are based on the gradient descent algorithm to minimize the prediction error. This process is repeated many times, and the network continues to improve its predictions until the magnitude of the gradient is less than 1e-5.

Once trained, the network can be applied to new patients for whom the endpoint is unknown. The crucial parameters of a NN are the size-parameter (number of units in the layer) and the decay-parameter that penalizes large weights in the model to avoid overfitting (default=0). The modelling was done using the *nnet* function in the *nnet* library.

## Endpoint and predictor sets

With each modelling technique, we developed prediction models for 6-month mortality. We hereto used three predictor sets of increasing complexity, referred to as the Core, Core+CT and Core+CT+Lab sets (Table 2).

**Table 2** Admission characteristics and predictor sets

| **Admission characteristics** | | **Predictor set** | | |
|---|---|---|---|---|
| **Variable** | **Type** | **Core** | **Core+CT** | **Core+CT+Lab** |
| Age | continuous | x | x | x |
| Motor score | factor (4) | x | x | x |
| Pupils | factor (3) | x | x | x |
| Hypoxia | factor (2) | | x | x |
| Hypotension | factor (2) | | x | x |
| CT classification | factor (3) | | x | x |
| tSAH on CT | factor (2) | | x | x |
| Epidural mass on CT | factor (2) | | x | x |
| Glucose | continuous | | | x |
| Hb | continuous | | | x |
| ( )=number of categories | | | | |

When predictor values were missing, we used a single imputation technique to fill in missing values, based on correlations between predictor variables. We recognize that single imputation underestimated the variability of the parameter estimates in the model, but was expected to have only minor impact on the point estimates in the model compared to more complex procedures such as multiple imputation [23]. The point estimates are most relevant for prediction in future patients.

## Validation procedure

A modelling technique was applied to each of the fifteen data sets, which served as a development set for a prediction model. The performance of the resulting model was calculated for the development set ("apparent performance"). The model was then ap-

plied to each of the fourteen remaining data sets that each in turn served as validation set ("validated performance").
The development and validation process was repeated for each of the fifteen development sets, leading to a total of 15 x 14=210 validations per predictor set. With three predictor sets, we had 3 x 210 = 630 validations for a stable impression of comparative performance of the five modelling techniques.

### Performance

The performance of the prediction models was quantified by discrimination and calibration. The area under the receiver operator characteristic curve (AUC) indicated the discriminative ability of the models.
The apparent AUC value was calculated by developing a model on the development set by straightforward calculation of the AUC. The validated AUC value was calculated by applying the model to the validation set followed by straightforward calculation of the AUC [24].
The Cox recalibration framework was used to assess calibration of the prediction models [25]. We estimated the calibration slope as the logistic regression coefficient in a model for 6-month mortality that included the log odds of the predictions as the single predictor: mortality ~ logit($\hat{y}$), where $\hat{y}$ is the predicted probability of mortality. We recalibrated the models for the development sets so that the calibration slope is 1 when the apparent performance is estimated in the development set. The slope is commonly smaller than 1 when validated in independent data [26].
Besides the area under the receiver operator characteristic curve (AUC) as measure for discriminative performance, the Brier score was calculated. This score is based on squared distances between predicted and observed outcomes. We scaled the Brier score such as to indicate the performance of a model with predictors against a non-informative model without predictors on a scale from 0 – 100% [27].

### Sensitivity analyses

Two sensitivity analyses were performed. The first analysis was performed with two non-linear models, RF and NN, as underlying model with the Core+CT+Lab predictor set. The second analysis considered the variables "Motor score", "Pupils" and "CT-class" as continuous variables instead of categorical variables (Appendix 2).

## 7.3 RESULTS

For each model and for each predictor set, the median apparent AUC values were over 0.7 for each of the three predictor sets with each of the five techniques, except for the

**Table 3** Median apparent and validated AUC values over analyses in 15 data sets

| Type | Model | LR | CART | RF | SVM | NN |
|---|---|---|---|---|---|---|
| Apparent | Core | 0.738 | 0.708 | 0.676 | 0.717 | 0.763 |
| | Core+CT | 0.794 | 0.733 | 0.718 | 0.801 | 0.821 |
| | Core+CT+Lab | 0.812 | 0.744 | 0.750 | 0.833 | 0.878 |
| Validated | Core | 0.725 | 0.654 | 0.690 | 0.664 | 0.710 |
| | Core+CT | 0.764 | 0.669 | 0.730 | 0.728 | 0.706 |
| | Core+CT+Lab | 0.757 | 0.666 | 0.735 | 0.732 | 0.674 |

RF model with the Core predictor set (0.676, Table 3). Validated AUC values were lower than at development, with largest optimism for the most complex predictor set. Specifically, apparent performance was optimistic for the SVM and NN models (Apparent AUC 0.833 and 0.878 versus validated AUC 0.732 and 0.674 respectively). The CART models showed relatively poor validated performance (AUC <0.7). The LR models showed the best validated AUC values closely followed by the RF models. The median of the validated calibration slopes of the NN models was remarkably low for the Core+CT and the Core+CT+Lab predictor sets (0.45 and 0.32 respectively, Table 4). The LR models showed the highest median validated scaled Brier scores (Table 5, scaled Brier > 10%), while RF and NN models predicted worse than chance (negative scaled Brier scores).

Figure 2 shows the variability of the validated AUC values of the models for each development set and for each of the three predictor sets. The LR and RF models showed a reasonably stable performance for each of the development sets, while performance varied more with other modelling techniques. The sensitivity analysis with respect to other underlying models showed that the LR models still had the highest AUC values.

**Table 4** Median validated calibration slopes over analyses in 15 data sets

| Model | LR | CART | RF | SVM | NN |
|---|---|---|---|---|---|
| Core | 0.92 | 0.68 | 0.93 | 0.76 | 0.67 |
| Core+CT | 0.80 | 0.65 | 1.03 | 0.69 | 0.45 |
| Core+CT+Lab | 0.78 | 0.54 | 1.03 | 0.63 | 0.32 |

**Table 5** Median validated scaled Brier scores over analyses in 15 data sets

| Model | LR | CART | RF | SVM | NN |
|---|---|---|---|---|---|
| Core | 10% | 4% | −14% | 3% | 6% |
| Core+CT | 14% | 3% | 6% | 8% | −8% |
| Core+CT+Lab | 13% | −2% | 9% | 6% | −14% |

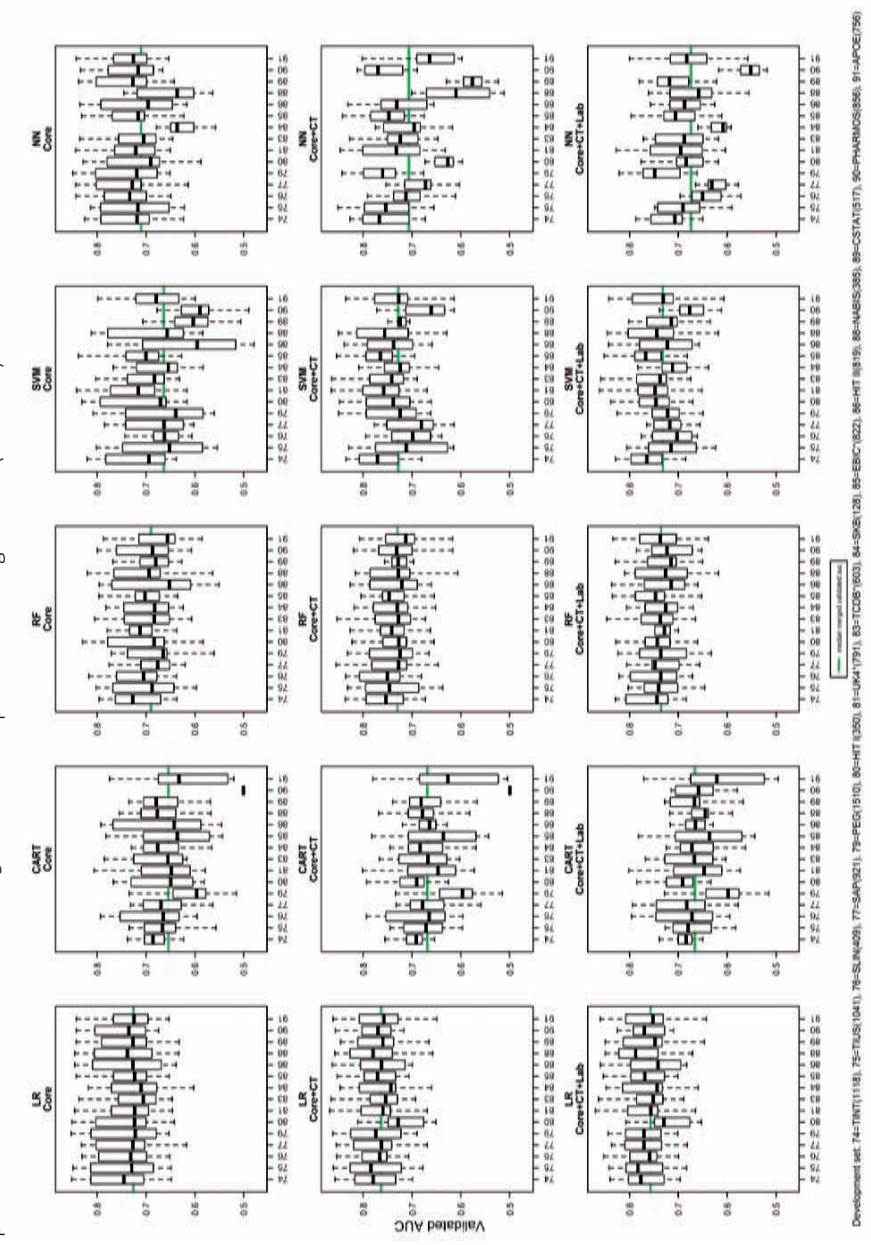



**Figure 2** Boxplots of validated AUC values across 15 data sets with patients suffering from TBI (n=11026)

Using continuous rather than factor coding for the variables "Motor score", "Pupils" and "CT-class" had no impact on the results (Appendix 2).

## 7.4 DISCUSSION

This systematic comparison of the external validity of prediction models revealed wide differences in performance between logistic regression models and four other modelling techniques in predicting 6-month mortality of patients suffering from traumatic brain injury (TBI). The classic logistic regression (LR) models performed best, closely followed by the random forest (RF) and support vector machine (SVM) models. The models based on classification and regression techniques (CART) and neural net (NN) showed a disappointing performance. These findings were consistent over three sets of predictors of increasing complexity, although SVM appeared to perform relatively better with a more complex predictor set.

In addition to the average discriminative power, stability of the performance at external validation is also important. The LR and RF models achieved quite stable validated AUC values with each predictor set. The stable performance of the RF models over the validation sets might be explained by the fact that the RF modelling technique internally validates multiple models using bootstrapping. The SVM models only showed stable performance for the most complex predictor set. This might be explained by the higher dimensional setting of this set. Substantial variation in validated performance was found for models based on NN and CART.

An important role of prediction models is to inform patients on their prognosis. A natural requirement to a model is that predictions are well calibrated [10]. For a fair comparison, we recalibrated the models at the development sets so that the calibration slope for each modelling technique equals one. The validated calibration of the RF models was best, followed by the LR models. The NN models showed poor calibration at validation.

For better insight in the relative performance of the models, we also calculated the scaled Brier score for each data set [10]. Comparison of model performance with these scaled Brier scores revealed that LR modelling, with each predictor set, clearly outperformed the other approaches.

While the interest in the development, validation, and clinical application of clinical prediction models is increasing, a recent systematic review showed that only a quarter of the studies reported prediction models with internal as well as external validation [27] [28]. Examples of internal validation techniques are split-sample, cross-validation and bootstrapping [29]. External validation, on the other hand, aims to address the

performance of a model in patients from a different but plausibly related setting, which still represents the underlying disease domain. This validation step is widely considered necessary before implementing a developed prediction model in clinical practice [30] [31]. Our study supports this notion, specifically for models developed with modern techniques.

We found that NN models had a considerably lower validated performance than LR models at external validation for all fifteen cohorts considered. No development cohort could be found in which NN models systematically outperformed simple, classical LR models. This is in contrast with the findings of two previous studies, where NN models performed equally well as or better than LR models in predicting the 6-month mortality of patients after severe brain injury [32] [33]. In both studies, the results were based on some form of internal validation. This illustrates that an honest internal validation may be difficult to perform, and show overoptimistic results compared to external validation [30].

A strength of our study was the adequate sample size in each of the fifteen different data sets in the IMPACT database. The median sample size of 791 patients, with a median of 188 events, implies more than 10 events per variable also in the most complex model considered (Appendix 1, Table 1). This sample size should provide sufficient opportunity to reliably model non-linear and non-additive effects, if present. The real world character of our study also contributes to its value to compare the validity of modern modelling techniques, in contrast to simulation studies.

Our findings support the suggestion of earlier studies that RF modelling is attractive in medical decision problems. Note that large databases need to be available for model development, because RF is a "data hungry" modelling technique [34] [35]. We expect that more flexible models such as RF will outperform LR only with much larger sample sizes, in line with a previous study where flexible models such as GAM, MARS, NN and CART did not perform better than LR models when 25,000 patients from the GUSTO-I data set were used for model development [36].

RF modelling was also found attractive compared to NN, SVM, CART and three traditional classifiers, such as discriminant analysis and logistic regression, to predict dementia and cognitive impairment using 10 neuropsychological predictors with a data set of 400 subjects [6]. The authors concluded that RF and linear discriminant analysis had the highest discriminative power. While this study performed no external validation, internal validation was reported using 5-fold cross validation.

While RF and LR models consistently performed well in the present study, CART models consistently performed poorly. The poor performance of CART modelling may be explained by the fact that this technique categorizes continuous variables, with optimal cut-offs determined from all possible cut-off points. In the present study, the variable 'age' already was a very strong predictor when analyzed continuously [37],

and therefore categorization would lead to substantial information loss. Furthermore, the CART modelling technique assumes higher-order interactions between all predictor variables, which may be unnecessary and harmful for discriminatory ability. RF modelling is an obvious improvement over CART modelling [38]. It is hence remarkable that CART has until recently been advocated as the preferred modelling technique for prediction in some disease areas, such as trauma [4].

Our findings on CART modelling contrast with a study that compared CART and NN with logistic regression in predicting cardiovascular risk. This study concluded that CART and NN can complement existing models [5]. The discrepancy between this study and ours may partly be explained again by the fact that only internal validation was attempted by splitting the dataset into a training set (n=10,296) and a test set (n=5,148). External validation was not performed. In addition, sample size may have been sufficient to allow for reliable estimation of the of CART and NN models.

Our sensitivity analysis by simulating dichotomous outcomes from two non-linear underlying models, RF and NN, with the Core+CT+Lab predictor set revealed that LR modelling still outperformed the other modelling techniques. This might be explained by the lack of strong non-linear effects in the data.

One of the limitations of our study was the use of default settings for the modelling techniques. This holds for LR, where penalization with ridge regression, or lasso type of methods might be used, as well as for the modern methods, where various specifics might be fine-tuned to the development setting [1] [3] [38]. Further tuning of parameters to specific issues in a particular development data set might obviously improve apparent performance, but we doubt that substantial improvement would be achieved in the validated external performance.

We conclude that in the area of predicting mortality from traumatic brain injury, non-linear and non-additive effects are not pronounced enough to make modern prediction methods beneficial. Random forests and support vector machines may lead to similar performance as logistic regression if very large data sets are available with larger sets of potential predictors. Neural networks and in particular CART may be harmful for prediction. Since performance may vary substantially across settings, external validation is a necessary step before applying a prediction model in a new setting, specifically if a modern technique was used for model development.

## 7.5 ACKNOWLEDGEMENTS

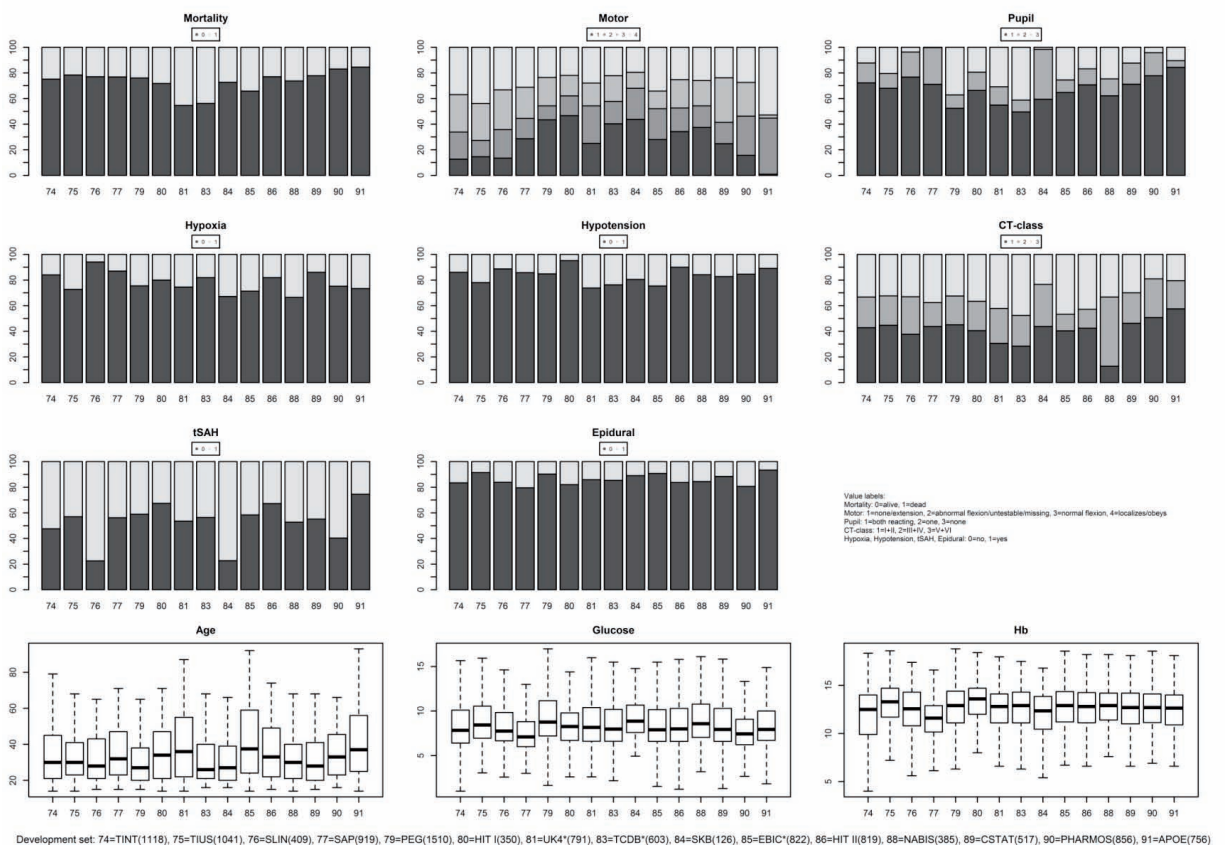
The authors thank the IMPACT researchers group for providing the data, Dr. B. van Calster for methodological support and Mrs. L. van Hulst for editorial support.

## 7.6 APPENDIX 1 CHARACTERISTICS DATA SETS

**Table 1** Characteristics data sets

| Study | Name | Period | Type | n | Mortality |
|---|---|---|---|---|---|
| 1 | TINT | 1991-1994 | RCT | 1118 | 25% |
| 2 | TIUS | 1991-1994 | RCT | 1041 | 22% |
| 3 | SLIN | 1994-1996 | RCT | 409 | 23% |
| 4 | SAP | 1995-1997 | RCT | 919 | 23% |
| 5 | PEG | 1993-1995 | RCT | 1510 | 24% |
| 6 | HIT I | 1987-1989 | RCT | 350 | 28% |
| 7 | UK4 | 1986-1988 | OBS | 791 | 45% |
| 8 | TCDB | 1984-1987 | OBS | 603 | 44% |
| 9 | SKB | 1996-1996 | RCT | 126 | 27% |
| 10 | EBIC | 1995-1995 | OBS | 822 | 34% |
| 11 | HIT II | 1989-1991 | RCT | 819 | 23% |
| 12 | NABIS | 1994-1998 | RCT | 385 | 26% |
| 13 | CSTAT | 1996-1997 | RCT | 517 | 22% |
| 14 | PHARMOS | 2001-2004 | RCT | 856 | 17% |
| 15 | APOE | 1996-1999 | OBS | 756 | 15% |

**Figure 1** Characteristics 15 data sets with patients suffering from TBI (n=11026)



## 7.7 APPENDIX 2 SENSITIVITY ANALYSES

This appendix comprises the results of simulations with two non-linear underlying models (RF and NN) with the Core+CT+Lab predictor set (Table 6, 7) and the results with the variables "Motor score", "Pupils" and "CT-class" considered as continuous variables instead of categorical variables (Table 8, 9, 10).

Results simulations with two non-linear underlying models with the Core+CT+Lab predictor set over analyses in 15 data sets

**Table 6** Medians of validated values with RF as underlying model

| | LR | CART | RF | SVM | NN |
|---|---|---|---|---|---|
| AUC | 0.675 | 0.585 | 0.650 | 0.628 | 0.586 |
| Scaled Brier score | 1.9% | −17.5% | 1.3% | −9.3% | −27.3% |
| Slope | 0.621 | 0.293 | 0.819 | 0.291 | 0.154 |

**Table 7** Medians of validated values with NN as underlying model

| | LR | CART | RF | SVM | NN |
|---|---|---|---|---|---|
| AUC | 0.726 | 0.612 | 0.704 | 0.711 | 0.641 |
| Scaled Brier score | 5.5% | −16.5% | 4.4% | −0.6% | −19.6% |
| Slope | 0.617 | 0.399 | 0.672 | 0.476 | 0.221 |

Results with the variables "Motor score", "Pupils" and "CT-class" considered as continuous variables instead of categorical variables over analyses in 15 data sets

**Table 8** Median apparent and validated AUC values

| | Model | LR | CART | RF | SVM | NN |
|---|---|---|---|---|---|---|
| Apparent | Core | 0.738 | 0.708 | 0.681 | 0.708 | 0.758 |
| | Core+CT | 0.787 | 0.733 | 0.733 | 0.809 | 0.824 |
| | Core+CT+Lab | 0.809 | 0.749 | 0.744 | 0.848 | 0.843 |
| Validated | Core | 0.734 | 0.655 | 0.691 | 0.660 | 0.712 |
| | Core+CT | 0.768 | 0.669 | 0.737 | 0.724 | 0.720 |
| | Core+CT+Lab | 0.762 | 0.664 | 0.736 | 0.727 | 0.702 |

**Table 9** Median validated calibration slopes

| Model | LR | CART | RF | SVM | NN |
|---|---|---|---|---|---|
| Core | 0.919 | 0.695 | 0.956 | 0.764 | 0.681 |
| Core+CT | 0.859 | 0.661 | 1.005 | 0.724 | 0.447 |
| Core+CT+Lab | 0.806 | 0.541 | 1.012 | 0.613 | 0.451 |

**Table 10** Median validated scaled Brier scores

| Model | LR | CART | RF | SVM | NN |
|---|---|---|---|---|---|
| Core | 10.7% | 4.1% | −13.9% | 3.9% | 6.5% |
| Core+CT | 15.3% | 3.5% | 6.9% | 9.3% | −1.1% |
| Core+CT+Lab | 13.9% | −0.9% | 9.6% | 7.4% | −0.7% |

## 7.8 APPENDIX 3 R-CODE MODELLING AND VALIDATION

This appendix contains the R-code with the development and validation of prediction models for 6-month mortality with the Core+CT+Lab predictor set in TBI data sets.

**General part**

```
# start
setwd("D:\\")
rm(list=ls(all=TRUE))

# open libraries
library(foreign)
library(caret)
library(rms)
library(rpart)
library(e1071)
library(randomForest)
library(caTools)
library(nnet)

# read SPSS file
D<- read.spss("priority100709finalImputed6.sav", use.value.labels=F, max.value.labels=Inf,
    to.data.frame=TRUE)

# set level of measurement
D$trial<-as.factor(D$trial)
D$d_motor<-as.factor(D$d_motor)
D$i_pupil<-as.factor(D$i_pupil)
D$d_mort<-as.factor(D$d_mort)
D$i_hypoxa<-as.factor(D$i_hypoxa)
D$i_hypots<-as.factor(D$i_hypots)
D$i_ctclas<-as.factor(D$i_ctclas)
```

```
D$i_tsah<-as.factor(D$i_tsah)

D$i_edh<-as.factor(D$i_edh)

D$i_glucos<-as.numeric(D$i_glucos)

D$i_hb<-as.numeric(D$i_hb)

D$i_age<-as.numeric(D$age)
```

**# define dataset with Core+CT+Lab predictors and mortality**

```
vars <- c("d_mort", "trial", "age", "d_motor", "i_pupil","i_hypoxa","i_hypots","i_ctclas","i_tsah","i_
        edh","i_glucos","i_hb")

D<-D[vars]
```

**Model-specific part**

**# loops for development and validation LR**

```
output<-matrix(ncol=11,nrow=210)

k=1

index<-c(74,75,76,77,79,80,81,83,84,85,86,88,89,90,91)

for( i in index)

{

DEV<-D[D$trial==i,]

set.seed(1)

Model<-glm(as.factor(d_mort)      ~age+d_motor+i_pupil+i_hypoxa+i_hypots+i_ctclas+i_tsah+i_
        edh+i_glucos+i_hb, data = DEV,family="binomial")

pdev<-predict(Model,type="response")

prev<-mean(as.numeric(DEV$d_mort)-1)

briermaxdev<-prev*(1-prev)

chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)

a<-chardev[12]

b<-chardev[13]

pdev<-plogis(a+b*qlogis(pdev))

chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)

index2<-index[index!=i]

for( j in index2)
```

```
{
VAL<-D[D$trial==j,]
pval<-predict(Model,VAL,type="response")
prev<-mean(as.numeric(VAL$d_mort)-1)
briermaxval<-prev*(1-prev)
pval<-plogis(a+b*qlogis(pval))
charval<-val.prob(pval,as.numeric(VAL$d_mort)-1,VAL,pl=F)
auccmc<-AUCcmc(pval)
output[k,]<-c(i,j,1-chardev[11]/briermaxdev,chardev[2],chardev[12],chardev[13],1-charval[11]/brier
    maxval,charval[2],charval[12],charval[13],auccmc)
print(k)
k<-k+1
}
}
```

**# loops for development and validation CART**

```
output<-matrix(ncol=11,nrow=210)
k=1
index<-c(74,75,76,77,79,80,81,83,84,85,86,88,89,90,91)
for( i in index)
{
DEV<-D[D$trial==i,]
set.seed(1)
Model<-rpart(as.factor(d_mort)      ~age+d_motor+i_pupil+i_hypoxa+i_hypots+i_ctclas+i_tsah+i_
    edh+i_glucos+i_hb, data = DEV)
pdev<-predict(Model)[,c("1")]
prev<-mean(as.numeric(DEV$d_mort)-1)
briermaxdev<-prev*(1-prev)
chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)
a<-chardev[12]
b<-chardev[13]
pdev<-plogis(a+b*qlogis(pdev))
```

```
chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)

index2<-index[index!=i]

for( j in index2)

{

VAL<-D[D$trial==j,]

pval<-predict(Model,VAL,type="prob")[,c("1")]

prev<-mean(as.numeric(VAL$d_mort)-1)

briermaxval<-prev*(1-prev)

pval<-plogis(a+b*qlogis(pval))

charval<-val.prob(pval,as.numeric(VAL$d_mort)-1,VAL,pl=F)

auccmc<-AUCcmc(pval)

output[k,]<-c(i,j,1-chardev[11]/briermaxdev,chardev[2],chardev[12],chardev[13],1-charval[11]/brier
        maxval,charval[2],charval[12],charval[13],auccmc)

print(k)

k<-k+1

}

}
```

**# loops for development and validation RF**

```
output<-matrix(ncol=11,nrow=210)

k=1

index<-c(74,75,76,77,79,80,81,83,84,85,86,88,89,90,91)

for( i in index)

{

DEV<-D[D$trial==i,]

set.seed(1)

Model<-randomForest(as.factor(d_mort)   ~age+d_motor+i_pupil+i_hypoxa+i_hypots+i_ctclas+i_
        tsah+i_edh+i_glucos+i_hb, data = DEV)

pdev<-predict(Model,type="prob")[,c("1")]

prev<-mean(as.numeric(DEV$d_mort)-1)

briermaxdev<-prev*(1-prev)

chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)
```

```
a<-chardev[12]
b<-chardev[13]
pdev<-plogis(a+b*qlogis(pdev))
chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)
index2<-index[index!=i]
for( j in index2)
{
VAL<-D[D$trial==j,]
pval<-predict(Model,VAL,type="prob")[,c("1")]
prev<-mean(as.numeric(VAL$d_mort)-1)
briermaxval<-prev*(1-prev)
charval<-val.prob(pval,as.numeric(VAL$d_mort)-1,VAL,pl=F)
pval<-plogis(a+b*qlogis(pval))
charval<-val.prob(pval,as.numeric(VAL$d_mort)-1,VAL,pl=F)
auccmc<-AUCcmc(pval)
output[k,]<-c(i,j,1-chardev[11]/briermaxdev,chardev[2],chardev[12],chardev[13],1-charval[11]/brier
        maxval,charval[2],charval[12],charval[13],auccmc)
print(k)
k<-k+1
}
}
```

**# loops for development and validation SVM**

```
output<-matrix(ncol=11,nrow=210)
k=1
index<-c(74,75,76,77,79,80,81,83,84,85,86,88,89,90,91)
for( i in index)
{
DEV<-D[D$trial==i,]
set.seed(1)
Model <-svm(as.factor(d_mort) ~age+d_motor+i_pupil+i_hypoxa+i_hypots+i_ctclas+i_tsah+i_
        edh+i_glucos+i_hb, data = DEV,probability=T)
```

```
pdev<-predict(Model,DEV,probability=T)
prev<-mean(as.numeric(DEV$d_mort)-1)
briermaxdev<-prev*(1-prev)
pdev<-attributes(pdev)$probabilities[,c("1")]
chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)
a<-chardev[12]
b<-chardev[13]
pdev<-plogis(a+b*qlogis(pdev))
chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)
index2<-index[index!=i]
for( j in index2)
{
VAL<-D[D$trial==j,]
pval<-predict(Model,VAL,probability=T)
pval<-attributes(pval)$probabilities[,c("1")]
prev<-mean(as.numeric(VAL$d_mort)-1)
briermaxval<-prev*(1-prev)
charval<-val.prob(pval,as.numeric(VAL$d_mort)-1,VAL,pl=F)
pval<-plogis(a+b*qlogis(pval))
charval<-val.prob(pval,as.numeric(VAL$d_mort)-1,VAL,pl=F)
auccmc<-AUCcmc(pval)
output[k,]<-c(i,j,1-chardev[11]/briermaxdev,chardev[2],chardev[12],chardev[13],1-charval[11]/brier
      maxval,charval[2],charval[12],charval[13],auccmc)
print(k)
k<-k+1
}
}
```

**# loops for development and validation NN**

```
output<-matrix(ncol=11,nrow=210)
k=1
```

```
index<-c(74,75,76,77,79,80,81,83,84,85,86,88,89,90,91)
for( i in index)
{
DEV<-D[D$trial==i,]
set.seed(1)
Model<-nnet(as.factor(d_mort)      ~age+d_motor+i_pupil+i_hypoxa+i_hypots+i_ctclas+i_tsah+i_
      edh+i_glucos+i_hb, data = DEV,size=10)
pdev<-predict(Model)
prev<-mean(as.numeric(DEV$d_mort)-1)
briermaxdev<-prev*(1-prev)
chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)
a<-chardev[12]
b<-chardev[13]
pdev<-plogis(a+b*qlogis(pdev))
chardev<-val.prob(pdev,as.numeric(DEV$d_mort)-1,DEV,pl=F)
index2<-index[index!=i]
for( j in index2)
{
VAL<-D[D$trial==j,]
pval<-predict(Model,VAL)
prev<-mean(as.numeric(VAL$d_mort)-1)
briermaxval<-prev*(1-prev)
charval<-val.prob(pval,as.numeric(VAL$d_mort)-1,VAL,pl=F)
pval<-plogis(a+b*qlogis(pval))
charval<-val.prob(pval,as.numeric(VAL$d_mort)-1,VAL,pl=F)
auccmc<-AUCcmc(pval)
output[k,]<-c(i,j,1-chardev[11]/briermaxdev,chardev[2],chardev[12],chardev[13],1-charval[11]/br
      iermaxval,charval[2],charval[12],charval[13],auccmc)
print(k)
k<-k+1
}
}
```

## REFERENCES


1. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R: The Elements of Statistical Learning. Volume 2. Springer; 2009.
2. Harrell FE: Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer; 2001.
3. Steyerberg EW: Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Volume 19; 2009.
4. Young NH, Andrews PJD: Developing a Prognostic Model for Traumatic Brain Injury—A Missed Opportunity? PLoS Med 2008, 5:3.
5. Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent MC: Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. In Proc AMIA Symp; 2000:156.
6. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A: Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes 2011, 4:299.
7. Terrin N, Schmid CH, Griffith JL, D'Agostino RB, Selker HP: External validity of predictive models: A comparison of logistic regression, classification trees, and neural networks. J Clin Epidemiol 2003, 56:721–729.
8. Ecke TH, Hallmann S, Koch S, Ruttloff J, Cammann H, Gerullis H, Miller K, Stephan C: External validation of an artificial neural network and two nomograms for prostate cancer detection. ISRN Urol 2011, 2012:643181.
9. König IR, Malley JD, Weimar C, Diener H-C, Ziegler A: Practical experiences on the necessity of external validation. Stat Med 2007, 26:5499–5511.
10. 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 traditional and novel measures. Epidemiology 2010, 21:128–138.
11. Austin PC, Steyerberg EW: Predictive accuracy of risk factors and markers: A simulation study of the effect of novel markers on different performance measures for logistic regression models. Stat Med 2013, 32:661–672.
12. Vergouwe Y, Moons KGM, Steyerberg EW: External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010, 172:971–980.
13. Lingsma HF, Roozenbeek B, Steyerberg EW, Murray GD, Maas AI: Early prognosis in traumatic brain injury: from prophecies to predictions. Lancet Neurol 2010:543–554.
14. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JDF, Maas AIR: Predicting outcome after traumatic brain injury:

Development and international validation of prognostic scores based on admission characteristics. PLoS Med 2008, 5:1251–1261.

15. Jennett B, Teasdale G, Braakman R, Minderhoud J, Knill-Jones R: Predicting outcome in individual patients after severe head injury. Lancet 1976, 307:1031–1034.

16. Mushkudiani NA, Hukkelhoven CWPM, Hernández A V, Murray GD, Choi SC, Maas AIR, Steyerberg EW: A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol 2008, 61:331–343.

17. Cortes C, Vapnik V: Support-vector networks. Mach Learn 1995, 20:273–297.

18. Marmarou A, Lu J, Butcher I, McHugh GS, Mushkudiani NA, Murray GD, Steyerberg EW, Maas AIR: IMPACT database of traumatic brain injury: design and description. J Neurotrauma 2007, 24:239–250.

19. Roozenbeek B, Chiu Y-L, Lingsma HF, Gerber LM, Steyerberg EW, Ghajar J, Maas AIR: Predicting 14-day mortality after severe traumatic brain injury: application of the IMPACT models in the brain trauma foundation TBI-trac® New York State database. J Neurotrauma 2012, 29:1306–12.

20. Tufféry S: Data Mining and Statistics for Decision Making. John Wiley & Sons; 2011.

21. Breiman L, Friedman JH, Olshen RA, Stone CJ: Classification and Regression Trees. Volume 19; 1984.

22. R Development Core Team R: R: A Language and Environment for Statistical Computing. R Found Stat Comput 2011:409.

23. Maas AIR, Marmarou A, Murray GD, Teasdale SGM, Steyerberg EW: Prognosis and clinical trial design in traumatic brain injury: the IMPACT study. J Neurotrauma 2007, 24:232–238.

24. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982, 143:29–36.

25. Cox DR: Two Further Applications of a Model for Binary Regression. Biometrika 1958, 45:pp. 562–565.

26. van Houwelingen JC, Le Cessie S: Predictive value of statistical models. Stat Med 1990, 9: 1303–1325.

27. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu L-M, Moons KG, others: External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014, 14:40.

28. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA: External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015, 68:25–34.

29. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF: Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001, 54:774–781.

30. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, Moons KGM: External validation is necessary in prediction research: A clinical example. J Clin Epidemiol 2003, 56:826–832.

31. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Ann Intern Med 1999, 130:515–524.

32. Lang EW, Pitts LH, Damron SL, Rutledge R: Outcome after severe head injury: an analysis of prediction based upon comparison of neural network versus logistic regression analysis. Neurol Res 1997, 19:274–280.

33. Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E: Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak 2005, 5:3.

34. van der Ploeg T, Datema F, de Jong RB, Steyerberg EW: Prediction of Survival with Alternative Modelling Techniques Using Pseudo Values. PLoS One 2014, 9:e100234.

35. van der Ploeg T, Austin PC, Steyerberg EW: Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014, 14:137.

36. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R: A comparison of statistical learning methods on the Gusto database. Stat Med 1998, 17:2501–2508.

37. Hukkelhoven C, Steyerberg E, Habbema J, Farace E, Marmarou A, Murray G, Marshall L, Maas A: Predicting Outcome after Traumatic Brain Injury: Admission Characteristics. J Neurotrauma 2005, 22:1025–39.

38. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD: Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000, 19:1059–1079.

# Chapter 8

## Assessing discriminative performance at external validation of clinical prediction models

D Nieboer, T van der Ploeg, EW Steyerberg

PLoS One. 2016 Feb 16;11(2):e0148820. doi: 10.1371/journal.pone.0148820.

## ABSTRACT

### Introduction

External validation studies are essential to study the generalizability of prediction models. Recently a permutation test, focusing on discrimination as quantified by the *c*-statistic, was proposed to judge whether a prediction model is transportable to a new setting. We aimed to evaluate this test and compare it to previously proposed procedures to judge any changes in *c*-statistic from development to external validation setting.

### Methods

We compared the use of the permutation test to the use of benchmark values of the c-statistic following from a previously proposed framework to judge transportability of a prediction model. In a simulation study we developed prediction mosel with logistic regression on a development set and validated them in the validation set. We concentrated on two scenarios: 1) the case-mix was more heterogeneous and predictor effects were weaker in the validation set compared to the development set, and 2) the case-mix was less heterogeneous in the validation set and predictor effects were identical in the validation and development set. Furthermore we illustrated the methods in a case study using 15 datasets of patients suffering from traumatic brain injury.

### Results

The permutation test indicated that the validation and development set were homogenous in scenario 1 (in almost all simulated samples) and heterogeneous in scenario 2 (in 17%-39% of simulated samples). Previously proposed benchmark values of the c-statistic and the standard deviation of the linear predictors correctly pointed at the more heterogeneous case-mix in scenario 1 and the less heterogeneous case-mix in scenario 2.

### Conclusion

The recently proposed permutation test may provide misleading results when externally validating prediction models in the presence of case-mix differences between the development and validation population. To correctly interpret the *c*-statistic found at external validation it is crucial to disentangle case-mix differences from incorrect regression coefficients.

## 8.1 INTRODUCTION

Clinical prediction models receive increasing attention for medical practice and research. After the development of a prediction model, an external validation study is essential to explore whether predictions done by the model are valid in a new population (1,2). The discriminative ability of prediction models is often quantified using a concordance (*c*) statistic (3). The *c*-statistic measures whether a prediction model can discriminate between patients with and without the outcome of interest. For logistic regression models the *c*-statistic is equivalent to the area under the receiver operating characteristic (ROC) curve (AUC) (4).

External validation studies are considered the stronger tests for a model compared to internal validation procedures such as cross-validation or bootstrap resampling (5). The possible differences between the validation and development setting make an external validation study a test of 'transportability' of a prediction model (6). If the validation population contains similar patients as the development population, the external validation study could merely be considered a test of the 'reproducibility' of a prediction model. Reproducibility refers to the ability of a prediction model to give valid predictions in a population very similar to the development population, whilst transportability refers to the ability to give valid predictions in populations that are related to but different from the development population (6). Typical examples of tests of transportability are assessment of model performance across different geographical regions or in different time periods.

We previously proposed a framework to identify if an external validation study investigated the reproducibility or transportability of a prediction model (7). This framework consists of three steps, 1) investigate the relatedness of the development and validation population, 2) validation of the prediction model in the new population, and 3) interpreting the results found at step 2 using the results from step 1. To assess the relatedness between the development and validation sample we proposed a membership model, i.e. a model predicting whether a patient is from the development or validation sample. Moreover, we suggested to compare the standard deviation of the linear predictors between development and validation samples, where the linear predictor is the linear combination of the regression coefficients from the model and the covariate values in the development and validation samples respectively.

Recently a permutation test was introduced with the aim to consolidate step 1 & 2 into a single step (8). The permutation test tests the hypothesis that the development and validation population are homogeneous. The permutation test obtains a p-value by judging the change in *c*-statistics of the model between the development and validation samples. The permutation test assesses the degree of homology between the development and validation sets. When the permutation test gives a p-value below

a pre-specified threshold, typically 0.05, the hypothesis that the development and validation samples are homogeneous is rejected. The claim was that the model may then not be directly transported to the validation population without further revision or updating (8).

Previous research has shown that the *c*-statistic does not only depend upon the validity of the prediction model, i.e. correctness of the regression coefficients, but also on the case-mix, i.e. the heterogeneity between patients in the population (9). A measure of the case-mix heterogeneity is the standard deviation of the linear predictor in a sample. Benchmark values of the *c*-statistic have also been developed to disentangle case-mix effects from the effects of incorrect regression coefficients (10). One such benchmark value is called the model based *c*-statistic (mbc). This is the expected *c*-statistic in a population given that the predictions made by a model are perfectly valid.

We aimed to evaluate the usefulness of the recently proposed permutation test in relation to the previously proposed framework. Specifically, we compare the conclusions from this test with conclusions drawn using earlier proposed measures, i.e. the standard deviation of the linear predictor and benchmark values of the *c*-statistic. We evaluated the different measures using a simulation study and in a case study using 15 datasets containing patients suffering from traumatic brain injury (TBI).

## 8.2 METHODS

We considered three strategies to judge a change in *c*-statistic when externally validating a prediction model: a permutation test; the standard deviation of the linear predictor; and benchmark values of the *c*-statistic.

### Simulation study

To assess the performance of the proposed permutation test we conducted a simulation study. In the simulation we varied the case-mix differences and predictor effects between the development and validation population, we also varied the sample sizes available for development and validation. In our simulation study we generated a development set **D** and validation set **V** using the following model:

$$y_{ij} \sim \text{bernoulli}(\pi_{ij}),$$
$$\pi_{ij} = \text{logit}^{-1}(\beta_j x_{ij}),$$
$$x_{ij} \sim N(0, \sigma_j^2),$$

Where $i$ denotes the patient number and $j \in \{D,V\}$ indicates whether the patient belongs to the development or validation set.

We distinguished three scenarios. In the first scenario we assumed that the case-mix distribution and predictor effects in the development and validation dataset are homogeneous. In the second situation we assumed that the case-mix is more heterogeneous and predictor effects were weaker in the validation set compared to the development set. In the third situation we assumed that the case-mix was less heterogeneous in the validation set as compared to the development set, but that the predictor effects were similar. The different values of the parameters in the different situations are shown in table 1.

**Table 1:** Parameters of the different scenarios in the simulation study

| | **SD predictor development population ($\sigma_D$)** | **SD predictor validation population ($\sigma_V$)** | **Coefficient development population ($\beta_D$)** | **Coefficient validation population ($\beta_V$)** |
|---|---|---|---|---|
| Homogeneous populations | 1 | 1 | 3 | 3 |
| Different case-mix & predictor effects | 1 | 1.5 | 3 | 2 |
| Different case-mix & same predictor effects | 1 | 0.75 | 3 | 3 |

SD: standard deviation

For each situation we generated development and validation sets containing 40, 100 or 200 patients, resulting in 9 scenarios in total. We developed a prediction model in the development set using logistic regression and validated the resulting model in the validation set. Subsequently we calculated the standard deviation of the linear predictor in the development and validation sets, the ratio of the standard deviation of the linear predictor in the validation and development set, two benchmark values for the *c*-statistic and performed the permutation test. R-scripts used for the simulation study are available as supplementary material (Appendix).

## Case study

Our case study uses data from the IMPACT database (11). This database contains data from 15 studies with patients suffering from traumatic brain injury (TBI). This database was previously used to develop a prediction model predicting 6-month mortality and unfavorable outcome using patient characteristics such as age, motor score and pupillary reactivity (12). In our case study we developed a model predicting 6-month mortality using the on the international arm of the Tirilizad trial (13). The prediction model contained the predictors age, motor score and pupillary reactivity. We subsequently

validated the developed model in the remaining 14 studies and judged the change in *c*-statistic observed at development and at external validation using the the ratio of the standard deviation of the linear predictor in the validation and development set, benchmark values of the *c*-statistic, and the permutation test. All statistical analyses were done in R 3.1.2 (14).

### Permutation test

The permutation test was developed to test the null hypothesis that the development and validation populations are homogeneous. When the nullhypothesis is not rejected it should be safe to transport the prediction model from the development to the validation population. We then claim that the model is valid in the validation population. The permutation test starts with calculating the observed *c*-statistic at external validation, denoted by $c^V$, of a prediction model. The *c*-statistic is calculated by comparing predictions from the model to observed outcomes. Subsequently patients are randomly permuted between the development and validation population. A prediction model is developed on the permuted development set and the *c*-statistic of this model is estimated in the permuted validation population. This process is repeated *k* times. The p-value of the permutation test is given by the proportion of times that the *c*-statistic of the model developed on the permuted development set was smaller than $c^V$. Whenever the p-value is below a prespecified threshold, typically 0.05, the null hypothesis is rejected that the development and validation population are homogeneous. The prediction model should then be updated before being transported to the validation population. For our simulation and case study we used a value of *k* equal to 1,000.

### Measures of case-mix

A direct way to investigate the difference in case-mix between the development and validation population is to compare the standard deviation (SD) of the linear predictor of the prediction model in the development and validation population. The linear predictor is the linear combination of the regression coefficients from the model and the covariate values in the development and validation samples respectively:
$lp_D = X_D\beta_D$ ; $lp_V = X_V\beta_D$.
A population with a more heterogeneous case-mix has a higher SD of the linear predictor compared to a more homogeneous case-mix.

### Benchmark values

The discriminative ability of a prediction model at external validation can be influenced by both the correctness of regression coefficients and the case-mix heterogeneity in the validation sample. This was a key point in our proposed framework and other work (7,10). Since differences in case-mix have no impact on the validity of the prediction

model it is important to distinguish between the influence of incorrect regression coefficients and case-mix. Therefore, two benchmark values of the *c*-statistic were proposed, the model based *c*-statistic (mbc) and the *c*-statistic obtained by refitting the model in the validation sample ($c^{\text{refitted}}$) (10). The mbc is the expected *c*-statistic in a population given that the prediction model is correct. Differences between the mbc and observed c-statistic at external validation ($c^{V}$) indicate the extent of poor model fit independent of differences in case-mix between development and validation samples. The mbc can be obtained by first calculating the predicted probability for each patient in the validation sample and subsequently generating a new outcome value based on this probability (10). To ensure stable estimates of the benchmark values at least 100 repetitions for each subject are required. The refitted *c*-statistic ($c^{\text{refitted}}$) gives an upper bound on the performance of the model in the validation population, if the regression coefficients from the prediction model are perfectly valid. Comparison of $c^{V}$ and $c^{\text{refitted}}$ reflects the influence of incorrect regression coefficients, given a similar case-mix as in the validation population. The mbc uses the regression coefficients from the prediction model developed in the development sample, while $c^{\text{refitted}}$ uses regression coefficients from the validation sample. Interpretation of $c^{V}$ is possible by considering the combination of the validity of the regression coefficients (as learned from comparison to $c^{\text{refitted}}$) and the case-mix (difference to development sample learned from comparison to mbc).

## 8.3 RESULTS

### Simulation study

In the scenario where the development and validation population were homogeneous, and the development and validation sets were relatively small, the median $c^{D}$ and $c^{V}$ were both 0.92 (Table 2).

Benchmark values of the *c*-statistic indicated that the regression coefficients of the model and case-mix between the development and validation sets were similar. Since the *c*-statistic is a rank based measure, by definition $c^{V}$ and $c^{\text{refitted}}$ were equal to each other in our simulations. The standard deviation of the linear predictor was similar in the development and validation sets, correctly indicating that the case-mix was similar in both samples, the median ratio of both standard deviations was close to 1. The permutation test rejected the null hypothesis of homogeneity between the validation and development sets in approximately 5% of the generated samples in the simulation study. Changing the sample size of the development and validation sets yielded similar results, however the interquartile range became somewhat smaller.

When both the case-mix and regression coefficients in the development and validation population were different the median $c^{D}$ and $c^{V}$ were both equal to 0.92. The median

**Table 2:** Results from the simulation study, median and inter-quartile range of 1000 simulations is shown for c-statistics and standard deviations. Proportion of samples where the permutation test rejected the null hypothesis is shown.

| | $c^D$ | $c^V$ | SD $lp_D$ | SD $lp_V$ | Ratio SD $lp_V$ and SD $lp_D$ | mbc | $c^{refitted}$ | Proportion samples null-hypothesis rejected |
|---|---|---|---|---|---|---|---|---|
| **Same case-mix and predictor effects** | | | | | | | | |
| Small sample (epv: 20) | 0.92 (0.89-0.95) | 0.92 (0.90-0.95) | 3.14 (2.52-4.30) | 3.16 (2.57-4.14) | 1.01 (0.91-1.12) | 0.93 (0.90-0.95) | 0.92 (0.90-0.95) | 0.06 |
| Medium sample size (epv: 50) | 0.92 (0.90-0.94) | 0.92 (0.90-0.94) | 3.08 (2.63-3.52) | 3.05 (2.65-3.55) | 1.00 (0.93-1.07) | 0.92 (0.90-0.94) | 0.92 (0.90-0.94) | 0.05 |
| Large sample size (epv: 100) | 0.92 (0.91-0.93) | 0.92 (0.91-0.93) | 3.04 (2.75-3.37) | 3.04 (2.74-3.39) | 0.99 (0.95-1.05) | 0.92 (0.91-0.93) | 0.92 (0.91-0.93) | 0.06 |
| **Different case-mix and predictor effects** | | | | | | | | |
| Small sample (epv: 20) | 0.92 (0.89-0.95) | 0.92 (0.89-0.95) | 3.23 (2.58-4.14) | 4.85 (3.89-6.22) | 1.49 (1.35-1.67) | 0.96 (0.95-0.98) | 0.92 (0.90-0.95) | 0.03 |
| Medium sample size (epv: 50) | 0.92 (0.90-0.94) | 0.92 (0.90-0.94) | 2.06 (2.67-3.67) | 4.56 (4.00-5.32) | 1.50 (1.41-1.60) | 0.96 (0.95-0.97) | 0.92 (0.90-0.94) | 0.01 |
| Large sample size (epv: 100) | 0.92 (0.91-0.93) | 0.92 (0.91-0.94) | 3.04 (2.75-3.37) | 4.56 (4.12-5.06) | 1.50 (1.43-1.57) | 0.96 (0.95-0.97) | 0.92 (0.91-0.94) | 0.02 |
| **Different case-mix and same predictor effexts** | | | | | | | | |
| Small sample (epv: 20) | 0.92 (0.89-0.95) | 0.88 (0.85-0.92) | 3.22 (2.55-4.23) | 2.42 (1.90-3.18) | 0.75 (0.67-0.84) | 0.89 (0.86-0.93) | 0.88 (0.85-0.92) | 0.18 |
| Medium sample size (epv: 50) | 0.92 (0.90-0.94) | 0.88 (0.86-0.90) | 3.08 (2.63-3.61) | 2.30 (2.01-2.70) | 0.75 (0.70-0.80) | 0.89 (0.86-0.91) | 0.88 (0.86-0.90) | 0.25 |
| Large sample size (epv: 100) | 0.92 (0.91-0.93) | 0.88 (0.86-0.90) | 3.03 (2.74-3.37) | 2.27 (2.06-2.52) | 0.75 (0.71-0.79) | 0.88 (0.87-0.90) | 0.88 (0.86-0.90) | 0.39 |

$c^D$: *c*-statistic of the model in the development population
$c^V$: *c*-statistic at external validation
SD $lp_D$: standard deviation linear predictor in development population
SD $lp_V$: standard deviation linear predictor in validation population
mbc: model based *c*-statistic
$c^{refitted}$: *c*-statistic of model refitted in the validation population

mbc was equal to 0.96, by definition $c^{refitted}$ was equal to 0.92. This indicated that although $c^{V}$ did not change, a substantially higher *c*-statistic was expected if the regression coefficients of the original model had been correct. With small sample sizes, the median standard deviation of the linear predictor was 3.23 in the development set and somewhat larger (4.85) in the validation set, indicating a more heterogeneous case-mix. The median ratio of the standard deviation of the linear predictor in the validation and development sample also indicated that the case mix was more heterogeneous in the validation population (1.50). The permutation test rejected the null hypothesis in 3% of the generated samples. Results using larger sample sizes were similar, however the interquartile range became somewhat smaller.

In the third scenario, the case-mix in the validation population was less heterogeneous compared to the development population. Using small sample sizes for model development and validation, the median $c^{D}$ was 0.92 and the median $c^{V}$ was 0.88. The median mbc was equal to 0.89 and median $c^{refitted}$ was 0.88, indicating that the drop in *c*-statistic between the development and validation set was due to a less heterogeneous case-mix in the validation set rather than incorrect regression coefficients. The median standard deviation of the linear predictor in the development population was 3.22 and somewhat smaller (2.42) in the validation set, indicating that the case-mix distribution was less heterogeneous in the validation set, which was confirmed by the median ratio of the standard deviations (0.75). The permutation test rejected the null hypothesis of homogeneous population in approximately 18% of the cases. Increasing the available sample sizes of the development and validation sets showed similar results, except for the permutation test where the proportion of samples where the null hypothesis was rejected increased to 39% as the available sample size increased, reflecting more statistical power. Again the inter-quartile range became somewhat smaller.

## Case study

The prediction model developed in the international arm of the Tirilizad trial had a *c*-statistic of 0.71 [0.67-0.74 95%CI]. The standard deviation of the linear predictor at development was equal to 0.80. When the model was externally validated the *c*-statistic ranged between 0.64 and 0.85 (Table 3).

If the standard deviation of the linear predictor was larger in the validation sample than in the development sample, then the model based *c*-statistic (mbc) was larger than the *c*-statistic at development. This reflected the wider spread of the risk distributions. The permutation test indicated evidence of heterogeneity between the development and validation sample in 4 out of 14 validations. However, the mbc and the standard deviation of the linear predictor indicate that the decrease in *c*-statistic in the SLIN study was mainly attributable to a less heterogeneous case-mix distribution, rather than incorrect regression coefficients. The permutation test indicated no evidence of

**Table 3:** External validation results of the model predicting 6-month mortality in TBI patients using age, motor score and pupillary reactivity.

| Study | $c^V$ | mbc | SD $lp_V$ | Ratio SD $lp_V$ and SD $lp_D$ | $c^{refittted}$ | p-value permutation test |
|---|---|---|---|---|---|---|
| TINT | 0.71[1] | - | - | - | - | - |
| TIUS | 0.74 | 0.73 | 0.87 | 1.09 | 0.74 | 1.00 |
| SLIN | 0.68 | 0.69 | 0.71 | 0.89 | 0.68 | 0.00 |
| SAP | 0.69 | 0.74 | 0.95 | 1.19 | 0.74 | 0.00 |
| PEG | 0.76 | 0.78 | 1.17 | 1.46 | 0.77 | 1.00 |
| HIT I | 0.72 | 0.77 | 1.12 | 1.40 | 0.79 | 0.90 |
| UK4 | 0.81 | 0.78 | 1.16 | 1.45 | 0.83 | 1.00 |
| TCDB | 0.82 | 0.80 | 1.25 | 1.56 | 0.83 | 1.00 |
| SKB | 0.68 | 0.75 | 1.01 | 1.26 | 0.72 | 0.35 |
| EBIC | 0.83 | 0.79 | 1.24 | 1.55 | 0.85 | 1.00 |
| HIT II | 0.69 | 0.77 | 1.10 | 1.38 | 0.73 | 0.00 |
| NABIS | 0.69 | 0.76 | 1.04 | 1.30 | 0.72 | 0.55 |
| CSTAT | 0.75 | 0.72 | 0.86 | 1.08 | 0.77 | 1.00 |
| PHARMOS | 0.64 | 0.70 | 0.76 | 0.95 | 0.66 | 0.00 |
| APOE | 0.85 | 0.73 | 0.86 | 1.08 | 0.85 | 1.00 |

[1] *c*-statistic of the model at development
$c^V$: *c*-statistic observed at external validation
mbc: model based *c*-statistic
SD $lp_V$: standard deviation of the linear predictor in the validation data
SD $lp_D$: standard deviation of the linear predictor in the development data
$c^{refitted}$: *c*-statistic of the prediction model refitted in the validation data

heterogeneity in 10 out of the 14 validation studies. Evaluation of the mbc and standard deviation of the linear predictor led to the same conclusion for the TIUS and PEG studies. In the other 8 cases there was a substantial influence of incorrect regression coefficients on the observed *c*-statistic at external validation, indicating that the model was not transportable to these settings.

## 8.4 DISCUSSION

This study illustrated how two separate phenomena determine differences in observed discriminative ability, i.e. the *c*-statistic, between development and validation settings. Case-mix and the correctness of regression coefficients both influence the *c*-statistic of a prediction model when applied in a new datasets. Attempts to provide a single summary test for differences in *c*-statistic are therefore misleading. The recently proposed permutation test incorrectly concluded that the development and validation population were not homogeneous in the scenario with different case-mix but similar

predictor effects. Conversely the permutation test concluded that development and validation population were homogeneous when the case-mix was more heterogeneous but predictor effects were weaker. Similar patterns were observed in the case study.

The permutation test aimed to consolidate the first two steps in the framework proposed by Debray et al. (7), by judging the heterogeneity between development and validation population using the change in *c*-statistic of the prediction model. The *c*-statistic however does not only depend on whether a prediction model gives valid predictions, but also on the case-mix in the underlying population; that was the key point in the framework by Debray. The permutation test does not take these case-mix differences into account and may break down when these are present.

When validating the IMPACT prediction model, predicting 6-month mortality of patients suffering from traumatic brain injury, it was noted that the *c*-statistic at external validation was higher in datasets from observational studies compared to the *c*-statistic found when validating in datasets from randomized controlled trials (15). These differences were attributed to the wider enrollment criteria in the observational studies compared to trials, leading to a more heterogeneous case-mix in the observational studies compared to the trials. Similarly, a recent review found higher *c*-statistic values in some validation studies than in the development studies, again suggesting that more heterogeneity at validation is well possible (16). The overall pattern in this review was a lower performance at validation than expected, reflecting overoptimism and overfitting at model development (4).

At external validation the performance of a prediction model is assessed using data not used at model development. Here we focused on judging the change in *c*-statistic of the prediction model at external validation. Validation studies however should also aim to assess other model properties, in particular the calibration of a prediction model. Calibration refers to the agreement between predicted probabilities and observed outcomes. It can adequately be assessed using recalibration parameters, and graphically using calibration plots (4,17).

The standard deviation of the linear predictor is a simple measure of case-mix heterogeneity in a dataset. When the distribution of the linear predictor is skewed the standard deviation may not be appropriate as a measure of case-mix heterogeneity. We note however that the distribution of the linear predictor is often close to a normal distribution (18). At external validation the distribution of the linear predictor should be assessed graphically in a 'validation' plot (19,20).In sum, the proposed permutation test does not take case-mix differences into account and can therefore give misleading results in the presence of case-mix differences. The permutation test therefore is only useful when there are no case-mix differences between the development and validation set. Case-mix differences between development and validation setting can readily be detected by simple summary measures such as the variance of the linear predictor

or benchmark values of the *c*-statistic. To judge the change in c-statistic of a prediction model at external validation it is crucial to disentangle the effects of incorrect regression coefficients from differences in case-mix heterogeneity between the development and validation setting.

## 8.5 APPENDIX

**R-script used for the simulation study**

```
rm(list = ls())
# Simulation study regarding the permutation test and other measures for judging
# the change in c-statistic
library(rms)

perm_auc <- function(D, V, k = 1000, AUCext){
# Permutation test
library(rms)
n_cases_D <- sum(D$y)
n_cases_V <- sum(V$y)

n_controls_D <- sum(1 - D$y)
n_controls_V <- sum(1 - V$y)

cases<- rbind(D[D$y==1, ], V[V$y==1, ])
controls <- rbind(D[D$y==0, ], V[V$y==0, ])

AUC <- rep(0, k)
index_D1 <- 1:n_cases_D
index_D0 <- 1:n_controls_D

index_V1 <- (n_cases_D + 1):(n_cases_D + n_cases_V)
index_V0 <- (n_controls_D + 1):(n_controls_D + n_controls_V)

for(j in 1:k){
cases_j<- cases[sample(1:nrow(cases)), ]
controls_j <- controls[sample(1:nrow(controls)), ]

D_j <- rbind(cases_j[index_D1, ], controls_j[index_D0, ])
```

```
V_j <- rbind(cases_j[index_V1, ], controls_j[index_V0, ])

fit_j <- lrm(y ~ x, data = D_j)
V_j$lp<- predict(fit_j, newdata = V_j)
fit_j_ext <- lrm(y ~ lp, data = V_j)
AUC[j]<- fit_j_ext$stats["C"]
}
p <- mean(AUC<AUCext, na.rm = T)
res <- list(p = p, AUC = AUC, AUC_ext = AUCext)
return(res)
}

mb.c <- function(xb.hat){
# Model based c-statistic
n<-length(xb.hat)
xb.hat <- rep(xb.hat, 100)
y<- plogis(xb.hat)>=runif(length(xb.hat))

library(pROC)
cstat <- roc(response = y, predictor = xb.hat)
mb.c<- auc(cstat)

return(list(mbc=mb.c))
}

n_sim <- 100
measures<- c('C_dev', 'C_ext', 'sd_dev', 'sd_val', 'mbc', 'ratio_sd',
'permutation', 'crefitted')

res <- array(dim =c(n_sim, length(measures)),
dimnames = list(n_sim = 1:n_sim,
```

```
measure = measures))
n<- 200 # Number of patients (note ~50% patients experience the event)
sd_d <- 1 # Standard deviation predictor in development population
sd_v <- 1 # Standard deviation predictor in validation population
beta_d <- 3 # Predictor effect development population
beta_v <- 3 # Predictor effect validation population

for(j in 1:n_sim){
Xd <- rnorm(n, sd = sd_d)
Xv <- rnorm(n, sd = sd_v)

yd <- plogis(beta_d * Xd)>=runif(n)
yv <- plogis(beta_v * Xv)>=runif(n)

D <- data.frame(cbind(Xd, yd))
names(D) <- c('x', 'y')

V <- data.frame(cbind(Xv, yv))
names(V) <- c('x', 'y')

fit_dev <- lrm(y ~ x, data = D)
sd_dev<- sd(fit_dev$linear.predictors)

lp_val<- cbind(1, V$x) %*% fit_dev$coefficients
fit_val <- lrm(V$y ~ lp_val)
refit <- lrm(y ~ x, data = V)

sd_val<- sd(lp_val)
p_value <- perm_auc(D = D, V = V, AUCext = fit_val$stats['C'])

res[j, 1] <- fit_dev$stats['C'] # Apparent c-statistic
```

```
res[j, 2] <- fit_val$stats['C'] # c-external
res[j, 3] <- sd_dev # Standard deviation of the linear predictor development
res[j, 4] <- sd_val # Standard deviation linear predictor at validation
res[j, 5] <- mb.c(lp_val)$mbc # model based c-statistic
res[j, 6] <- sd_val/sd_dev# Ratio standard deviation at validation and development
res[j, 7] <- p_value$p# p-value permutation test
res[j, 8] <- refit$stats['C'] # crefitted
}

apply(res, 2, quantile, probs = c(0.5, 0.25, 0.75))
```

## REFERENCES


1. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012; 98(9):691–8.
2. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med. 2013;10(2): e1001381.
3. Harrell FE, Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982 May 14;247(18):2543–6.
4. Steyerberg EW. Clinical Prediction Models. New York: Springer;
5. Bleeker S., Moll H., Steyerberg E., Donders AR., Derksen-Lubsen G, Grobbee D., et al. External validation is necessary in prediction research:: A clinical example. J Clin Epidemiol. 2003 Sep;56(9):826–32.
6. Justice AC, Covinsky KE, Berlin JA. Assessing the Generalizability of Prognostic Information. Ann Intern Med. 1999 Mar 16;130(6):515–24.
7. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015 Mar;68(3):279–89.
8. Wang L-Y, Lee W-C. A Permutation Method to Assess Heterogeneity in External Validation for Risk Prediction Models. PLoS ONE. 2015 Jan 21;10(1):e0116957.
9. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol. 2012;12.
10. Vergouwe Y, Moons KGM, Steyerberg EW. External Validity of Risk Models: Use of Benchmark Values to Disentangle a Case-Mix Effect From Incorrect Coefficients. Am J Epidemiol. 2010;172(8):971–80.
11. Maas AIR, Murray GD, Roozenbeek B, Lingsma HF, Butcher I, McHugh GS, et al. Advancing care for traumatic brain injury: findings from the IMPACT studies and perspectives on future research. Lancet Neurol. 2013 Dec;12(12):1200–10.
12. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting Outcome after Traumatic Brain Injury: Development and International Validation of Prognostic Scores Based on Admission Characteristics. PLoS Med. 2008 Aug 5;5(8):e165.
13. Hukkelhoven CWPM, Steyerberg EW, Habbema JDF, Farace E, Marmarou A, Murray GD, et al. Predicting Outcome after Traumatic Brain Injury: Development and Validation of a Prognostic Score Based on Admission Characteristics. J Neurotrauma. 2005 Oct 1;22(10): 1025–39.

14. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2008. Available from: http://www.R-project.org

15. Roozenbeek B, Lingsma HF, Lecky FE, Lu J, Weir J, Butcher I, et al. Prediction of Outcome after Moderate and Severe Traumatic Brain Injury: External Validation of the IMPACT and CRASH Prognostic Models. Crit Care Med. 2012 May;40(5):1609–17.

16. Siontis GCM, Tzoulaki I, Siontis KC, Ioannidis JPA. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ [Internet]. 2012 May 24; 344. Available from: http://www.bmj.com/content/344/bmj.e3318.abstract

17. Cox DR. Two Further Applications of a Model for Binary Regression. Biometrika. 1958 Dec 1;45(3/4):562–5.

18. Wessler BS, Lai YH L, Kramer W, Cangelosi M, Raman G, Lutz JS, et al. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. Circ Cardiovasc Qual Outcomes. 2015;8(4): 368–75.

19. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiol Camb Mass. 2010 Jan;21(1):128–38.

20. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33(3):517–35.

# Chapter 9

## General discussion

## GENERAL DISCUSSION

In this thesis, modern modelling techniques were compared with classical modelling techniques in a medical context. We evaluated to what extent modern modelling techniques have advantages in medical prediction problems: modern techniques are more flexible and may be better able to capture nonlinearity and interactions between predictor variables. They may therefore offer better opportunities for predicting medical outcomes, such as treatment success or mortality, for identifying risk factors to support diagnosis, and even for classifying sources of infections and contaminations to facilitate prevention. The aim of this research was to investigate in what circumstances and under what conditions relatively modern modelling techniques such as support vector machines, neural networks and random forests have advantages in medical prediction research over more classical modelling techniques, such as linear regression, logistic regression and Cox regression.

## PART I

### Answers to research questions

The thesis specifically focused on 3 research questions. The summary answers are shown in Table 1.

Question 1:
Comparison of modern and traditional modelling techniques:
- What is the performance in predicting intracranial findings on CT scans?
- What is the ability to capture nonlinearity?

Question 2:
Application of modern modelling techniques:
- How can they be applied for survival problems?
- How can they be applied for feature selection in a domain with many variables and comparatively few subjects or data points?

Question 3:
Performance of modern modelling techniques:
- What is the performance in relation to the sample size?
- What is the stability of the performance at external validation?

**Table 1** Summary questions and answers

| Question | Answer | Chapter |
|---|---|---|
| 1a Comparison of modern techniques in CT diagnosis | Performance of modern modelling techniques was not better than logistic regression. | 2 |
| 1b Ability to capture nonlinearity | Modern modelling techniques such as RCS and MFP are well suited to capture nonlinearity. | 3 |
| 2a Application of modern techniques in survival problems | Modern modelling techniques can be used straightforwardly to predict survival by using "pseudo values". | 4 |
| 2b Feature selection and prediction in "p>n" problems | Extensive bootstrapping led to models with a good performance based on a sparse feature set. | 5 |
| 3a Data hungriness of modern modelling techniques | Modern modelling techniques are far more data-hungry than traditional modelling techniques. | 6 |
| 3b Stability of the performance at external validation | Modern modelling techniques showed high optimism at external validation.<br>Permutation test for transportability may provide misleading results. | 7/8 |

### *1a Performance of modern modelling techniques in CT scans*

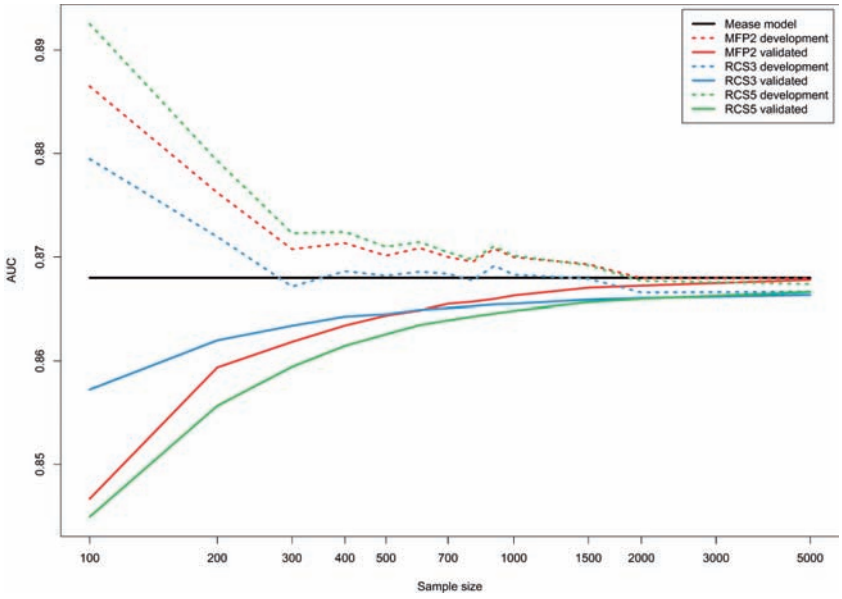
The performance of modern modelling techniques was assessed by comparing the performance of two classical modelling techniques (logistic regression and classification and regression trees) and five relatively modern modelling techniques (neural net, Bayes network, Chi square automatic interaction detection, decision list and support vector machines) in risk prediction on intracranial findings on CT-scans (Chapter 2). After 10x10 cross-validation, none of the included modern modelling techniques outperformed the logistic regression model, although Bayes network and neural net performed almost similarly. Besides, the Bayes network model has a presentation format that provides more detailed insights into the structure of the prediction problem, which might offer advantages in clinical practice. A particularly disappointing performance was shown by the support vector machine model. This model also showed a very high optimism, which was probably due to the used default setting of the parameters. This finding indicates that prediction modelling with the support vector machine technique implies the necessity of carefully tuning the parameters, without increasing the risk of overfitting.

### *1b Modern modelling techniques to capture nonlinearity*

The role of modern modelling techniques or machine learning techniques in medical settings was discussed in Chapter 3. Several predictive modelling issues were discussed, such as the role of tuning, model uncertainty and predictor selection. Simula-

tion studies showed that the modelling techniques multi fractional polynomials (MFP) and restricted cubic splines (RCS) both are well able to capture nonlinearity. These techniques are considered to be standard tools in regression modelling. The relation between the performance of the models generated with these techniques and sample size is reflected in Figure 1. The performance of MFP and RCS models was compared for increasing sample sizes (100 to 5000) based on simulated data (n=1,000,000) generated from the Mease model [1]. The MFP2 models have 4 degrees of freedom, 2 for finding the transformation and 2 for the coefficients. The RCS3 and RCS5 models have 2 and 4 degrees of freedom respectively [2]. Models were validated in a holdout sample with n=500,000. With increasing sample size, the optimism of the models decreased and the validated performance increased. For sample sizes =< 500, the validated performance of the RCS3 models was best and for sample sizes > 500, the validated performance of the MFP2 models was best. However, the differences were small from n>500 and negligible from n>2000 (Figure 1).

**Figure 1** AUC of RCS and MFP models in relation to the sample size




### *2a Modern modelling techniques for survival problems*

The use of modern modelling techniques in survival problems is hampered by the fact that these techniques are not suited for time-to-event outcomes (Chapter 4). Sometimes it is necessary to transform the predictor or outcome variables (e.g. log-transformation or categorization). In this case, the time-to-event outcome was transformed into new

single variables at certain time points, so-called pseudo values. These pseudo values were used to compare the performance of five modelling techniques. After bootstrap validation (200x), the best performance was achieved by classical modelling techniques. The logistic regression model was best for 60-month survival, and the general linear model was best for overall survival. However, the other models, including neural net, classification and regression trees and three variants of support vector machine models, performed almost equally well.

*2b Modern modelling techniques for feature selection with p>n*

The feature selection ability and validated predictive performance of modern modelling techniques was studied in the domain of legionella pneumophila. We evaluated many potential features (p) for classifying infections as clinical or environmental in a relatively small data set (p>n, Chapter 5). In such a situation, commonly used techniques are VARSEL-RF and SVM-RFE, which are based on stepwise elimination of irrelevant features. The result is a sparse set of important predictors. We showed that good prediction models could also be obtained by means of modern modelling techniques such as random forest, support vector machines and LASSO, which can first be used to preselect relevant features using bootstrapping and then to develop a prediction model with those selected features. The stability of the relevant feature sets was modest when determined by investigating the most important features of each bootstrap round. Our findings show that random forest and LASSO even performed slightly better than the commonly used techniques VARSEL-RF and SVM-RFE.

*3a Modern modelling techniques and data hungriness*

An important aspect in developing a prediction model is the required sample size of the development set. For logistic regression modelling, a common rule of thumb is that at least 10 events per variable (EPV) are required for sensible prediction modelling [3]. For other modelling techniques, such a rule of thumb is not available. The required sample size or "data hungriness" of modern modelling techniques was studied by sampling with increasing sample sizes from existing data sets (Chapter 6). Our analyses showed that modern modelling techniques such as support vector machines, neural networks and random forests may need over 10 times as high an EPV value (EPV>100) to achieve a stable AUC and a small optimism. This implies that such modern techniques should only be used in medical prediction problems if very large data sets are available. At present, this limits their usefulness in a medical context, since high-quality medical data sets are often small and have a poor signal-to-noise ratio. In the future, however, the access to, for example, routinely collected electronic health records (EHR) will offer opportunities for creating large databases with many patient records, and depending on the richness and quality of the data, modern modelling techniques may be applied.

*3b Performance at external validation*

The performance of five different modelling techniques at internal and external validation was investigated using fifteen data sets with TBI patients (Chapter 7). At external validation, the logistic regression models performed best, followed by random forest and support vector machine models. The CART models showed a poor performance at external validation, while these had been advocated specifically for this setting [4]. The most stable performance was achieved by the logistic regression models and the random forest models. Our study confirmed that external validation is an important step before applying a prediction model in a new setting. External validation aims to address the performance of a model in patients from a different but plausibly related setting that still represents the underlying disease domain. However, a data set with such patients is not always available, and in that case, internal validation is the only possibility. Again, the access to electronic heath records (EHR) may offer opportunities for creating validation data sets for external validation.

In Chapter 8, the transportability of a model was discussed based on a proposed permutation test for judging whether a prediction model is transportable to a new setting. The use of the permutation test was compared to the use of benchmark values of the refitted AUC and the model-based AUC. The conclusion was that the proposed test may provide misleading results in the presence of case mix differences between the development set and the validation set. The model-based AUC was also used in Chapter 7 when investigating the external validity of prediction models using fifteen data sets with TBI patients to gain insight into case mix differences.

*Strengths and weaknesses*

As mentioned above, an important aspect of various modern modelling techniques is the tuning of the hyper parameters to determine the best parameter setting based on the development set. In this thesis, a limitation was the use of default settings for the parameters as far as possible, except in predicting survival (Chapter 4). Although tuning of the hyper parameters is generally recommended, it must be kept in mind that this will generally lead to higher optimism. Therefore, validation of the optimized models is essential. Another limitation is that we did not test the modern modelling techniques in settings with very large data sets. In this thesis, most data sets had small to intermediate sizes.

One of the strengths of this thesis is that modern modelling techniques were compared in different medical settings with respect to different outcomes (dichotomous, continuous and time-to-event outcomes). Another strength is that thorough simulations were performed, especially for insight into the sample size that is required for a good and stable performance.

**Table 2** Strengths and weaknesses

| Aspect | Strength | Weakness |
|---|---|---|
| Modelling techniques | - Most common modelling techniques considered<br>- Different performance measures used | - Other techniques not considered<br>- Use of default settings for parameters instead of tuning |
| Medical settings | - Variety in settings<br>- Range of sample sizes<br>- Different types of outcomes | - Specific characteristics may have driven conclusions |
| Simulation sets | - True model known<br>- Sample size can be varied<br>- Full control over interactions | - Real world may be different<br>- Unrealistic functional form in medicine |

# PART II

## A perspective on prediction modelling

In this part of the discussion, various further aspects of predictive modelling will be discussed.

### *Understanding the modelling techniques*

In this thesis, various modelling techniques were applied and compared. A disadvantage of modern modelling techniques is that for most medical researchers the underlying algorithm is a black box. By contrast, classical modelling techniques using regression have the advantage that the resulting model is a relatively simple mathematical formula. The coefficients in this formula are calculated by optimizing the likelihood or the mean squared error, possibly with some form of penalization to optimize predictive performance. The coefficients can be used for instance to calculate odds ratios, hazard ratios, and rate ratios, all of which represent some form of relative risk. To fully understand the underlying algorithm of a modern modelling technique, a researcher needs in-depth mathematical and statistical knowledge. For the application of these modelling techniques, however, it may be sufficient to have some general knowledge about the working of the modelling techniques. In each chapter, we therefore described the modelling techniques used in this thesis at the level of knowledge required for applying these techniques (Chapter 2-8). This level includes knowledge of what kind of variables are permitted for the input and output, which model parameters are involved and what their function is, and how the output should be interpreted.

### *Tuning model parameters*

Some of the modern modelling techniques have so-called hyper parameters. Examples are the number of trees for random forest models, the tree-depth for tree models and the regularization parameter for support vector machine models. There are two pos-

sibilities to set the hyper parameters that are involved. One possibility is to use default settings, which are sometimes related to findings in the data. Another possibility is tuning the hyper parameters of a modelling technique using a grid search over supplied parameter ranges and using cross-validation to find the best parameter setting (Chapter 4). With the best parameter setting, a model can be developed on a development part of the data set, followed by validation of the model on a validation part.
To illustrate the effect of tuning the hyper parameters instead of using default settings, a sensitivity analysis (bootstrap resampling, n=200) was performed with three different support vector machine modelling techniques using the HNSCC data base. The outcome variable was dead or alive at 60 months and the eight predictor variables involved were Age, Gender, Tumor location, T-N-M classification, Prior malignancies and ACE27 as described in Chapter 4. Table 3 shows the mean AUC-values and calculated optimism as a result of this analysis. The validated AUC-values with tuning are slightly higher than the validated AUC-values with the default settings. However, optimism is also higher, meaning that tuning has its price.

**Table 3** Results sensitivity analysis

| | With tuning | | | With default settings | | |
|---|---|---|---|---|---|---|
| **Type SVM** | **AUC bootstrap** | **AUC validated** | **Optimism** | **AUC bootstrap** | **AUC validated** | **Optimism** |
| Linear | 0.805 | 0.794 | 0.011 | 0.805 | 0.794 | 0.011 |
| Polynomial | 0.862 | 0.812 | 0.051 | 0.792 | 0.774 | 0.018 |
| Radial | 0.871 | 0.812 | 0.059 | 0.820 | 0.801 | 0.019 |

*Performance measures*

To measure the performance of a model, we used the area under the receiver operator curve (AUC) to indicate the discriminatory ability of a model in case of a dichotomous outcome (Chapter 2-8), and in case of a continuous outcome we used the mean squared error (MSE) or the root of the mean squared error (RMSE) (Chapter 4). Calibration of the models was done using the Cox calibration framework, but only in Chapter 7. Another, less frequently used measurement for performance is the Brier score. We used a variant of this score, the scaled Brier score, for performance measurement in Chapter 7, for better interpretability.
It is difficult to define value judgments on what is a good or excellent performance. An interesting development is the focus on decision-analytic summary measures of model performance, specifically the calculation of Net Benefit (NB) [5]. Net Benefit is a simple type of decision analysis, with benefits and harms put on the same scale so that they can be compared directly [6]. Explaining such measures to clinicians is challenging. A

simple interpretation is that an NB higher than a reference of treat all or treat none implies that the model is useful.
Net benefit can also be expressed on a relative scale as Relative Utility. This scaling may be more attractive to some but not all researchers. Expression of NB on an absolute scale is consistent with decision making and cost-effectiveness, where the increase in effectiveness is balanced to the increase in costs.

*Physician versus prediction model*
In clinical practice, medical decision making is difficult. Prediction models may assist treating physicians to inform patients on the probability of an outcome, such as a 5-year disease-free period, or to classify a patient according to a certain risk [10]. The latter may also be important for communication between physicians and their patients. However, these calculations alone should never dictate patient care and are no substitute for professional judgement.

## Future developments

*Big Data*
The future of medical prediction research will be strongly influenced by the procurement and analysis of "Big Data". In 2013, IBM predicted that with the sharp increase of medical images and electronic medical records, medical professionals may utilize big data to extract useful clinical information from masses of data to obtain a medical history and forecast treatment effects, thus improving patient care and reducing cost [7]. Big data analysis is the process of examining large data sets containing a variety of data types to uncover hidden patterns, unknown correlations, trends and other useful information. Combining and analyzing electronic medical records, financial and operational data, clinical data and genomic data to match treatments with outcomes can help to predict patients at risk for disease or readmission and to provide more efficient care. Big data analytics and applications in healthcare are at a nascent stage of development, but their maturing process can be accelerated by rapid advances in platforms and tools ("Hadoop" for example) [8]. For predictive modelling with big data, the most frequently used modelling techniques are machine learning techniques, such as Bayesian networks [9]. However, this thesis shows that the advantages of machine learning techniques may be quite limited. Moreover, these techniques showed high risks of overfitting.

*Technological innovation*
Until fairly recently, applying relatively modern techniques for medical investigations was hampered by the fact that these techniques require in-depth statistical and

mathematical knowledge as well as knowledge in the field of informatics. Nowadays, however, these techniques are readily available in user-friendly programs that are incorporated in many software packages, such as R, SPSS modeler and Weka. Also, making the necessary calculations for the application of these techniques has become far less time-consuming due to the increased capacity and increased speed of modern computers.

New modelling techniques may be particularly useful for analyzing large databases with enormous numbers of patient records. Simulations can be performed to test the applicability and performance of different modelling techniques, in which predefined relations between predictor variables and outcome variables are used.

## Overall conclusion

In medical research and clinical practice, the modern modelling techniques evaluated in this thesis did not have important advantages over more classical modelling techniques, such as linear, logistic and Cox regression. These classical modelling techniques outperformed the relatively modern modelling techniques when predicting risk or predicting survival based on small- and medium-sized data sets. Since modern modelling techniques such as support vector machines, neural networks and random forests are more than ten times more data-hungry than regression techniques, a new rule of thumb for applying these techniques might be that the development set should contain at least 100 events per variable. If such large data sets are available, these techniques may have advantages.

## References


1. Mease D, Wyner A, Buja A: Boosted classification trees and class probability/quantile estimation. J Mach Learn Res 2007, 8:409–439.
2. Binder H, Sauerbrei W, Royston P: Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: A simulation study with continuous response. Stat Med 2013, 32:2262–2277.
3. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF: Prognostic modeling with logistic regression analysis in search of a sensible strategy in small data sets. Med Decis Mak 2001, 21:45–56.
4. Andrews PJD, Sleeman DH, Statham PFX, McQuatt A, Corruble V, Jones PA, Howells TP, Macmillan CSA: Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. J Neurosurg 2002, 97:326–336.
5. Vickers AJ, Elkin EB: Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak 2006, 26:565–574.
6. Vickers AJ, Van Calster B, Steyerberg EW: Net benefit approaches to the evaluation of prediction models, molecular markers and diagnostic tests. BMJ-British Med J 2015.
7. Chen M, Mao S, Liu Y: Big data: A survey. Mob Networks Appl 2014, 19:171–209.
8. Raghupathi W, Raghupathi V: Big data analytics in healthcare: promise and potential. Heal Inf Sci Syst 2014, 2:3.
9. Yoo C, Ramirez L, Liuzzi J: Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J 2014, 18:50–57.


.

# Chapter 10

## Miscellaneous

## SUMMARY

The aim of this research is to investigate in what circumstances and under what conditions relatively modern modelling techniques such as support vector machines, neural networks and random forests might have advantages in medical prediction research and clinical practice over more classical modelling techniques, such as linear regression, logistic regression and Cox regression.

One of the areas where modern modelling approaches have been advocated as an alternative to logistic regression is trauma research. Specifically, CART methods have been promoted. In Chapter 2 we investigated whether alternative modelling techniques might improve the performance of prediction rules for intracranial traumatic findings in patients with minor head injury. These prediction rules were designed to reduce the use of computed tomography (CT) without missing patients at risk for complications. We re-analyzed 3181 patients with minor head injury who had received CT scans between February 2002 and August 2004. Of these patients 243 (7.6%) had intracranial traumatic findings and 17 (0.5%) underwent neurosurgical intervention. We compared the sensitivity, specificity and area under the receiver operator curve of various modelling techniques and found that no modern modelling technique outperformed the logistic regression model. However, the Bayes network model had a presentation format which provided more detailed insights into the structure of the prediction problem.

The role of modern modelling techniques or machine learning techniques in medical settings was discussed in Chapter 3. Several predictive modelling issues such as tuning, model uncertainty and predictor selection were discussed. The discussion focused on the ability of machine learning techniques to capture nonlinearity in medical data. Simulation studies showed that modelling techniques such as fractional polynomials and restricted cubic splines are able to capture nonlinearity and that these techniques are considered as default tools in regression modelling.

In Chapter 4, we addressed the challenges that survival data pose to the application of modern modelling techniques. We used pseudo values to enable statistically appropriate analyses of survival outcomes predicted by means of seven alternative modelling techniques. We analyzed survival in 1282 Dutch patients with newly diagnosed Head and Neck Squamous Cell Carcinoma (HNSCC) with conventional Kaplan-Meier and Cox regression analysis. We subsequently calculated pseudo values to reflect the individual survival patterns. We used these pseudo values to compare the performance of models based on recursive partitioning (RPART), neural nets (NNET), logistic regression (LR), general linear models (GLM) and three variants of support vector machines (SVM)

with respect to dichotomous 60-month survival, and continuous pseudo values at 60 months or estimated survival time [13]. The best performance was achieved by the logistic regression model and the general linear model, followed by support vector machines with a linear kernel.

Chapter 5 addresses the genetic comparison of clinical and environmental *Legionella* strains, which forms an essential part of outbreak investigations. DNA microarrays often comprise many thousands of DNA markers (features). Feature selection and the development of prediction models are particularly challenging in this domain with many variables and comparatively few subjects or data points. We compared modelling techniques to develop prediction models for classifying legionella infections as clinical or environmental. We analyzed a database containing 222 Legionella pneumophila strains with 448 continuous markers and a dichotomous outcome (clinical or environmental) with four modelling techniques: classification and regression trees (CART), random forests (RF), support vector machines (SVM) and least absolute shrinkage and selection operator (LASSO). We found that in this domain, good prediction models were provided by the RF and LASSO techniques.

In Chapter 6, we studied the predictive performance of different modelling techniques in relation to the effective sample size. We labelled this relation "data hungriness". We performed simulation studies based on three clinical cohorts: 1282 patients with head and neck cancer (5-year survival 47%), 1731 patients with traumatic brain injury (6-month mortality 22.3%), and 3181 patients with minor head injury (7.6% with CT scan abnormalities). We compared three relatively modern modelling techniques: support vector machines (SVM), neural nets (NN) and random forests (RF), and two classical techniques: logistic regression (LR) and classification and regression trees (CART). Data hungriness was defined by plateauing of AUC and small optimism (difference between the mean apparent AUC and the mean validated AUC <0.01) [14]. The analysis showed that modern modelling techniques such as support vector machines, neural networks and random forests may need over 10 times as many events per variable to achieve a stable AUC and a small optimism as classical modelling techniques, such as logistic regression. This implies that such modern techniques should only be used in medical prediction problems if very large data sets are available.

In Chapter 7, we compared the externally validated performance of five prediction models for 6-month mortality of TBI patients with three predictor sets of increasing complexity. We used the IMPACT database on TBI patients with data of 15 underlying studies to compare the modelling techniques logistic regression (LR), classification and regression trees (CART), random forests (RF), support vector machines (SVM) and neural

nets (NN). External validation of the models was done by developing a model on one of the 15 data sets, followed by applying the model to each of the 14 remaining data sets. This process was repeated 15 times. The area under the receiver operator curve (AUC) was used to assess the performance of the models. The logistic regression models performed best, followed by random forest and support vector machine models. The CART models showed a poor performance. Our findings confirm that external validation is a necessary step before applying a prediction model in a new setting.

In Chapter 8, the transportability of a model was discussed based on a proposed permutation test for judging whether a prediction model is transportable to a new setting. The use of the permutation test was compared to the use of benchmark values of the refitted AUC and the model-based AUC. The conclusion was that the proposed test may provide misleading results in the presence of case mix differences between the development and validation set.

## SAMENVATTING

Het doel van dit onderzoek is te onderzoeken onder welke omstandigheden en onder welke condities relatief moderne modelleringstechnieken zoals support vector machines, neural networks en random forests voordelen zouden kunnen hebben in medisch-wetenschappelijk onderzoek en in de medische praktijk in vergelijking met meer traditionele modelleringstechnieken, zoals lineaire regressie, logistische regressie en Cox regressie.

Een van de gebieden waar moderne modelleringstechnieken werden aanbevolen als alternatief voor logistische regressie is trauma onderzoek. Met name CART werd aanbevolen als modelleringstechniek. In hoofdstuk 2 is onderzocht of moderne modelleringstechnieken de prestaties van voorspellingsregels voor intracraniële traumatische bevindingen bij patiënten met klein hoofdletsel zouden kunnen verbeteren. Deze voorspellingsregels werden ontworpen om het gebruik van computertomografie (CT) te verminderen zonder patiënten met risico op complicaties te missen. Er werden nieuwe analyses gemaakt van 3181 patiënten met klein hoofdletsel die tussen februari 2002 en augustus 2004 CT-scans hebben ondergaan. Van deze patiënten hadden 243 (7.6%) intracraniële traumatische bevindingen en 17 (0.5%) hadden een neurochirurgische ingreep ondergaan. Met maten als sensitiviteit, specificiteit en oppervlakte onder de receiver operator curve werden verschillende modelleringstechnieken vergeleken, met als uitkomst dat geen van de moderne modelleringstechnieken beter presteerde dan het logistische regressie model. Echter, het Bayes netwerkmodel had een presentatievorm die meer gedetailleerde inzichten biedt in de structuur van het predictieprobleem.

De rol van moderne modelleringstechnieken of machine learning technieken in medische omgevingen wordt besproken in hoofdstuk 3. Verschillende aspecten van het maken van predictiemodellen kwamen aan de orde, zoals tuning, model onzekerheid en de selectie van predictoren. De discussie richtte zich op het vermogen van machine learning technieken om met niet-lineariteit in medische gegevens om te gaan. Simulatiestudies toonden aan dat modelleringstechnieken zoals fractional polynomials (FP) en restricted cubic splines (RCS) niet-lineariteit kunnen modelleren en dat deze technieken kunnen worden beschouwd als standaard hulpmiddelen bij modellering met regressietechnieken.

In hoofdstuk 4 wordt ingegaan op de uitdagingen die het voorspellen van survival met zich meebrengt voor de toepassing van moderne modelleringstechnieken. We gebruikten zogenaamde pseudo-values om statistische analyses van survival mogelijk

te maken met zeven moderne modelleringstechnieken. De overleving van 1282 Nederlandse patiënten met recent gediagnosticeerd hoofd-hals carcinoom (in het Engels afgekort als HNSCC) werd geanalyseerd met conventionele Kaplan-Meier en Cox regressie analyse. Pseudo-values werden berekend om individuele overlevingspatronen van patiënten weer te geven. Met behulp van deze pseudo-values werden de prestaties vergeleken van modellen gebaseerd op recursieve partitioning (RPART), neural nets (NN), logistic regression (LR), general linear models (GLM) en drie varianten van support vector machines (SVM) ten aanzien van de 60-maands overleving, de pseudo-values bij 60 maanden en de geschatte overlevingstijd. Het LR-model en het GLM-model presteerden het best, gevolgd door het SVM-model met een lineaire kernel.

Hoofdstuk 5 behandelt de genetische vergelijking van klinische en omgeving gerelateerde legionella stammen, hetgeen een wezenlijk onderdeel vormt van onderzoeken van legionella-uitbraken. DNA microarrays bevatten vaak vele duizenden DNA-markers (variabelen). De selectie van relevante variabelen en de ontwikkeling van predictiemodellen is een probleem in dit domein met veel variabelen en relatief weinig gegevenspunten. Verschillende modelleringstechnieken voor het ontwikkelen van predictiemodellen werden vergeleken voor het categoriseren van legionella-infecties als klinisch of omgeving gerelateerd. We analyseerden een database met 222 Legionella pneumophila stammen met 448 permanente markers en een dichotome uitkomst (klinisch of omgeving gerelateerd) met behulp van vier modelleringstechnieken: classification and regression trees (CART), random forests (RF), support vector machines (SVM) en least absolute shrinkage and selection operator (LASSO). De technieken RF en LASSO leverden goede predictiemodellen op voor dit domein.

In hoofdstuk 6 worden de voorspellende prestaties van verschillende modelleringstechnieken in relatie tot de effectieve steekproefomvang bestudeerd. Deze relatie wordt de "data hungriness" genoemd. Simulatie studies werden uitgevoerd op basis van drie klinische cohorten: 1282 patiënten met hoofd-hals carcinoom (5-jaars overleving 47%), 1731 patiënten met traumatisch hersenletsel (6-maanden mortaliteit 22.3%) en 3181 patiënten met klein hoofdletsel (7.6 % met CT-scan afwijkingen). Drie relatief moderne modelleringstechnieken werden vergeleken: support vector machines (SVM), neural nets (NN) en random forests (RF) en twee klassieke technieken: logistische regressie (LR) en classification and regression trees (CART). De data hungriness werd vastgesteld op basis van de gevalideerde AUC en het optimisme van het model. Uit de analyse bleek dat moderne modelleringstechnieken zoals support vector machines, neural nets en random forests meer dan 10 keer zoveel "events" per variabele (EPV) nodig hebben als klassieke modelleringstechnieken zoals logistische regressie om een stabiele AUC en een laag optimisme te bereiken. Dit betekent dat dergelijke moderne

technieken alleen voor medische predictieproblemen kunnen worden gebruikt als er zeer grote datasets beschikbaar zijn.

In hoofdstuk 7 worden de extern gevalideerde prestaties van vijf predictiemodellen vergeleken voor de 6-maands mortaliteit van TBI patiënten op basis van drie sets met predictoren van toenemende complexiteit. We gebruikten de IMPACT-database met gegevens van 15 onderliggende studies om de modelleringstechnieken logistische regressie (LR), classification and regression trees (CART), random forests (RF), support vector machines (SVM) en neural nets (NN) te vergelijken. Externe validatie van de modellen werd uitgevoerd door een model te ontwikkelen op één van de 15 datasets, gevolgd door toepassing van het model op elk van de 14 resterende datasets. Deze werkwijze werd 15 keer herhaald. De oppervlakte onder de receiver operator curve (AUC) werd gebruikt om de prestaties van de modellen te kunnen beoordelen. De LR-modellen presteerden het best, gevolgd door de RF en SVM modellen. De CART modellen presteerden slecht. Onze bevindingen bevestigen dat externe validatie een noodzakelijke stap is voordat een predictiemodel in een nieuwe situatie wordt toegepast.

In hoofdstuk 8 wordt de transporteerbaarheid van een predictiemodel besproken op basis van een voorgestelde permutatietest die zou kunnen beoordelen of een predictiemodel naar een nieuwe situatie kan worden getransporteerd. Het gebruik van deze permutatietest werd vergeleken met het gebruik van benchmarkwaarden van verschillende typen AUC ("refitted" en "model-based"). De conclusie was dat de voorgestelde test misleidende resultaten kan opleveren indien er case-mix verschillen zijn tussen de data waarop het model wordt ontwikkeld en de data waarop het model wordt gevalideerd.

## LIST OF PUBLICATIONS


1. Al-Janabi S, van Slooten H-J, Visser M, van der Ploeg T, van Diest PJ, Jiwa M: Evaluation of mitotic activity index in breast cancer using whole slide digital images. PLoS One 2013, 8: e82576.
2. Apeldoorn AT, Bosselaar H, Ostelo RW, Blom-Luberti T, van der Ploeg T, Fritz JM, de Vet HCW, van Tulder MW: Identification of Patients With Chronic Low Back Pain Who Might Benefit From Additional Psychological Assessment. Clin J Pain 2012:23–31.
3. Apeldoorn AT, Ostelo RW, Fritz JM, van der Ploeg T, van Tulder MW, de Vet HCW: The Cross-sectional Construct Validity of the Waddell Score. Clin J Pain 2012:309–317.
4. Bolk J, van der Ploeg T, Cornel JH, Arnold AE, Sepers J, Umans VA: Impaired glucose metabolism predicts mortality after a myocardial infarction. Int J Cardiol 2001, 79:207–214.
5. Bouthoorn SH, van der Ploeg T, van Erkel NE, van der Lely N: Alcohol intoxication among Dutch adolescents: acute medical complications in the years 2000-2010. Clin Pediatr (Phila) 2011, 50:244–251.
6. Broers CJM, Sinclair N, van der Ploeg TJ, Jaarsma T, van Veldhuisen DJ, Umans VAWM: The post-infarction nurse practitioner project: A prospective study comparing nurse intervention with conventional care in a non-high-risk myocardial infarction population. Neth Heart J 2009, 17:61–67.
7. Broers CJM, Smulders J, van der Ploeg TJ, Arnold AER, Umans VAM: [Nurse practitioner equally as good as a resident in the treatment of stable patients after recent myocardial infarction, but with more patient satisfaction]. Ned Tijdschr Geneeskd 2006, 150:2544–2548.
8. Cakir H, van Stijn MFM, Lopes Cardozo AMF, Langenhorst BLAM, Schreurs WH, van der Ploeg TJ, Bemelman WA, Houdijk APJ: Adherence to Enhanced Recovery After Surgery and length of stay after colonic resection. Colorectal Dis 2013, 15:1019–1025.
9. Cakir H, Heus C, van der Ploeg TJ, Houdijk APJ: Visceral obesity determined by CT scan and outcomes after colorectal surgery; a systematic review and meta-analysis. Int J Colorectal Dis 2015, 30:875–882.
10. de Jonghe JFM, Kalisvaart KJ, Dijkstra M, van Dis H, Vreeswijk R, Kat MG, Eikelenboom P, van der Ploeg T, van Gool WA: Early symptoms in the prodromal phase of delirium: a prospective cohort study in elderly patients undergoing hip surgery. Am J Geriatr Psychiatry 2007, 15:112–121.
11. De Mulder M, Cornel JH, Van Der Ploeg T, Boersma E, Umans VA: Elevated admission glucose is associated with increased long-term mortality in myocardial infarction patients, irrespective of the initially applied reperfusion strategy. Am Heart J 2010, 160:412–419.
12. de Mulder M, van der Ploeg T, de Waard GA, Boersma E, Umans VA: Admission glucose does not improve GRACE score at 6 months and 5 years after myocardial infarction. Cardiology 2011, 120:227–234.

13. Dirkali A, van der Ploeg T, Nangrahary M, Cornel JH, Umans VAWM: The impact of admission plasma glucose on long-term mortality after STEMI and NSTEMI myocardial infarction. Int J Cardiol 2007:215–217.

14. Goede J, Hack WWM, Sijstermans K, van der Voort-Doedens LM, Van der Ploeg T, Meij-de Vries A, Delemarre-van de Waal HA: Normative values for testicular volume measured by ultrasonography in a normal population from infancy to adolescence. Horm Res Paediatr 2011, 76:56–64.

15. Goorden SMI, van Engelen RA, Wong LSM, van der Ploeg T, Verdel GJE, Buijs MM: A novel troponin I rule-out value below the upper reference limit for acute myocardial infarction. Heart 2016.

16. Kalisvaart KJ, Vreeswijk R, De Jonghe JFM, Van Der Ploeg T, Van Gool WA, Eikelenboom P: Risk factors and prediction of postoperative delirium in elderly hip-surgery patients: Implementation and validation of a medical risk factor model. J Am Geriatr Soc 2006, 54: 817–822.

17. Kat MG, de Jonghe JFM, Vreeswijk R, van der Ploeg T, van Gool WA, Eikelenboom P, Kalisvaart KJ: Mortality associated with delirium after hip-surgery: A 2-year follow-up study. Age Ageing 2011, 40:312–318.

18. Kat MG, Vreeswijk R, De Jonghe JFM, Van Der Ploeg T, Van Gool WA, Eikelenboom P, Kalisvaart KJ: Long-term cognitive outcome of delirium in elderly hip surgery patients: A prospective matched controlled study over two and a half years. Dement Geriatr Cogn Disord 2008, 26:1–8.

19. Kat MG, Zuidema SU, van der Ploeg T, Kalisvaart KJ, van Gool WA, Eikelenboom P, de Jonghe JFM: Reasons for psychiatric consultation referrals in Dutch nursing home patients with dementia: A comparison with normative data on prevalence of neuropsychiatric symptoms. Int J Geriatr Psychiatry 2008, 23:1014–1019.

20. Komen MMC, Breed WPM, Smorenburg CH, van der Ploeg T, Goey SH, van der Hoeven JJM, Nortier JWR, van den Hurk CJG: Results of 20- versus 45-min post-infusion scalp cooling time in the prevention of docetaxel-induced alopecia. Support care cancer Off J Multinatl Assoc Support Care Cancer 2016, 24:2735–2741.

21. Kramer GM, Leenders MWH, Schijf LJ, Go HLS, van der Ploeg T, van den Tol MP, Schreurs WH: Is ultrasound-guided fine-needle aspiration cytology of adequate value in detecting breast cancer patients with three or more positive axillary lymph nodes? Breast Cancer Res Treat 2016, 156:271–278.

22. Munsterman ID, Cleeren E, van der Ploeg T, Brohet R, van der Hulst R: "Pico-Bello-Klean study": effectiveness and patient tolerability of bowel preparation agents sodium picosulphate-magnesium citrate and polyethylene glycol before colonoscopy. A single-blinded randomized trial. Eur J Gastroenterol Hepatol 2015, 27:29–38.

23. Nieboer D, van der Ploeg T, Steyerberg EW: Assessing Discriminative Performance at External Validation of Clinical Prediction Models. PLoS One 2016, 11:e0148820.

24. Nielsen K, Poelman MM, den Bakker FM, van der Ploeg T, Bonjer HJ, Schreurs WH: Comparison of the Dutch and English versions of the Carolinas Comfort Scale: a specific quality-of-life questionnaire for abdominal hernia repairs with mesh. Hernia 2014, 18:459–464.

25. Nielsen K, Richir MC, Stolk TT, van der Ploeg T, Moormann GRHM, Wiarda BM, Schreurs WH: The limited role of ultrasound in the diagnostic process of colonic diverticulitis. World J Surg 2014, 38:1814–1818.

26. Oussoren E, Brands MMMG, Ruijter GJG, der Ploeg AT van, Reuser AJJ: Bone, joint and tooth development in mucopolysaccharidoses: relevance to therapeutic options. Biochim Biophys Acta 2011, 1812:1542–1556.

27. Paes BF, Vermeulen K, Brohet RM, van der Ploeg T, de Winter JP: Accuracy of tympanic and infrared skin thermometers in children. Arch Dis Child 2010, 95:974–978.

28. Poel YHM, Hummel P, Lips P, Stam F, van der Ploeg T, Simsek S: Vitamin D and gestational diabetes: a systematic review and meta-analysis. Eur J Intern Med 2012, 23:465–469.

29. Rood JAJ, van Zuuren FJ, Stam F, van der Ploeg T, Eeltink C, Verdonck-de Leeuw IM, Huijgens PC: Perceived need for information among patients with a haematological malignancy: associations with information satisfaction and treatment decision-making preferences. Hematol Oncol 2015, 33:85–98.

30. Rood JAJ, Van Zuuren FJ, Stam F, van der Ploeg T, Huijgens PC, Verdonck-de Leeuw IM: Cognitive coping style (monitoring and blunting) and the need for information, information satisfaction and shared decision making among patients with haematological malignancies. Psychooncology 2015, 24:564–571.

31. Slor CJ, Witlox J, Adamis D, Meagher DJ, van der Ploeg T, Jansen RWMM, van Stijn MFM, Houdijk APJ, van Gool WA, Eikelenboom P, de Jonghe JFM: Predicting delirium duration in elderly hip-surgery patients: does early symptom profile matter? Curr Gerontol Geriatr Res 2013, 2013:962321.

32. Snoek A, Dekker M, Lagrand T, Epema A, van der Ploeg T, van den Brand JGH: A clinical decision model identifies patients at risk for delayed diagnosed injuries after high-energy trauma. Eur J Emerg Med 2013, 20:167–172.

33. Steyerberg EW, van der Ploeg T, Van Calster B: Risk prediction with machine learning and regression methods. Biom J 2014, 00:1–6.

34. Taniguchi Y, Takahashi Y, Toba T, Yamada S, Yokoi K, Kobayashi S, Okajima S, Shimane A, Kawai H, Yasaka Y, Smanio P, Oliveira MA, Machado L, Cestari P, Medeiros E, Fukuzawa S, Okino S, Ikeda A, Maekawa J, Ichikawa S, Kuroiwa N, Yamanaka K, Igarashi A, Inagaki M, Patel K, Mahan M, Ananthasubramaniam K, Mouden M, Yokota S, Ottervanger JP, et al.: Poster Session 1: Sunday 3 May 2015, 08:30-18:00Room: Poster Area. Eur Heart J Cardiovasc Imaging 2015, 16 Suppl 1:i11–i28.

35. van Boven N, Theuns D, Bogaard K, Ruiter J, Kimman G, Berman L, VAN DER Ploeg T, Kardys I, Umans V: Atrial fibrillation in cardiac resynchronization therapy with a defibrillator: a risk factor for mortality, appropriate and inappropriate shocks. J Cardiovasc Electrophysiol 2013, 24:1116–1122.

36. van den Bergh B, Blankestijn J, van der Ploeg T, Tuinzing DB, Forouzanfar T: Conservative treatment of a mandibular condyle fracture: Comparing intermaxillary fixation with screws or arch bar. A randomised clinical trial. J Craniomaxillofac Surg 2015, 43:671–676.

37. van den Bergh B, de Mol van Otterloo JJ, van der Ploeg T, Tuinzing DB, Forouzanfar T: IMF-screws or arch bars as conservative treatment for mandibular condyle fractures: Quality of life aspects. J Craniomaxillofac Surg 2015, 43:1004–1009.

38. van den Bogaard VAB, Euser SM, van der Ploeg T, de Korte N, Sanders DGM, de Winter D, Vergroesen D, van Groningen K, de Winter P: Diagnosing perforated appendicitis in pediatric patients: a new model. J Pediatr Surg 2016, 51:444–448.

39. van der Ploeg T, Austin PC, Steyerberg EW: Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014, 14:137.

40. van der Ploeg T, Datema F, de Jong RB, Steyerberg EW: Prediction of Survival with Alternative Modeling Techniques Using Pseudo Values. PLoS One 2014, 9:e100234.

41. van der Ploeg T, Nieboer D, Steyerberg EW: Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. J Clin Epidemiol 2016.

42. van der Ploeg T, Smits M, Dippel DW, Hunink M, Steyerberg EW: Prediction of intracranial findings on CT-scans by alternative modelling techniques. BMC Med Res Methodol 2011: 143.

43. van der Ploeg T, Steyerberg EW: Feature selection and validated predictive performance in the domain of Legionella pneumophila: a comparative study. BMC Res Notes 2016, 9:147.

44. van Galen KPM, Visser HPJ, van der Ploeg T, Smorenburg CH: Prognostic factors in patients with breast cancer and malignant pleural effusion. Breast J 2010:675–677.

45. van Poppel MN, Koes BW, van der Ploeg T, Smid T, Bouter LM: Lumbar supports and education for the prevention of low back pain in industry: a randomized controlled trial. JAMA 1998, 279:1789–1794.

46. van Stijn MFM, Korkic-Halilovic I, Bakker MSM, van der Ploeg T, van Leeuwen PAM, Houdijk APJ: Preoperative nutrition status and postoperative outcome in elderly general surgery patients: a systematic review. JPEN J Parenter Enteral Nutr 2013, 37:37–43.

47. van Tellingen A, Grooteman MPC, Schoorl M, Bartels PCM, Schoorl M, van der Ploeg T, ter Wee PM, Nube MJ: Intercurrent clinical events are predictive of plasma C-reactive protein levels in hemodialysis patients. Kidney Int 2002, 62:632–638.

48. van Tellingen A, Grooteman MPC, Schoorl M, ter Wee PM, Bartels PCM, Schoorl M, van der Ploeg T, Nube MJ: Enhanced long-term reduction of plasma leptin concentrations by super-flux polysulfone dialysers. Nephrol Dial Transplant 2004, 19:1198–1203.

49. van Tellingen A, Schalkwijk CG, Teerlink T, Barto R, Grooteman MPC, van der Ploeg T, ter Wee PM, Nube MJ: Influence of different haemodialysis modalities on AGE peptide levels: intradialytic versus long-term results. Nephron Clin Pract 2005, 100:c1–7.

50. Van Zanten E, Van der Ploeg T, Van Hoof JJ, Van der Lely N: Gender, age, and educational level attribute to blood alcohol concentration in hospitalized intoxicated adolescents; A cohort study. Alcohol Clin Exp Res 2013, 37:1188–1194.

51. Wishaupt JO, van den Berg E, van Wijk T, van der Ploeg T, Versteegh F, Hartwig NG: Paediatric apnoeas are not related to a specific respiratory virus, and parental reports predict hospitalisation. Acta Paediatr 2016, 105:542–548.

52. Wondergem M, van der Zant FM, van der Ploeg T, Knol RJJ: A literature review of 18F-fluoride PET/CT and 18F-choline or 11C-choline PET/CT for detection of bone metastases in patients with prostate cancer. Nucl Med Commun 2013, 34:935–45.

## DANKWOORD

Allereerst wil ik mijn dank betuigen aan degenen die op enigerlei wijze een bijdrage hebben geleverd aan mijn proefschrift:

Prof.dr. Ewout Steyerberg, mijn promotor. Beste Ewout, dank voor de buitengewoon prettige wijze waarop je mij hebt begeleid en je deskundigheid hebt overgedragen. Je was altijd stimulerend op momenten waarop dat nodig was en daarnaast heb je ook een gevoel voor humor waardoor mijn vrijdag in Rotterdam een moment was om naar uit te kijken. Zonder jou was er nu geen proefschrift geweest.

Daan Nieboer, kamergenoot in de onderzoekerskamer op verdieping 24 van het Nagebouw van het Erasmus MC. Beste Daan, dank voor jouw waardevolle adviezen, jouw mooie verhalen en de mooie gesprekken over het voetbal. En die ene legendarische avond vergeet ik nooit.

Lisette van Hulst, mijn levensgezellin. Lieve Lisette, dank voor jouw hulp met het Engels en de adviezen met betrekking tot de structuur bij het schrijven van het proefschrift. Dank ook voor al je geduld en begrip als het tegen zat. Dank ook voor al die dingen die ik hier niet kan noemen.

Hogeschool Inholland, mijn werkgever. Dank voor de facilitering van twee dagen per week gedurende vier jaar om dit promotietraject mogelijk te maken.

De overige leden van de promotiecommissie, bestaande uit Dr.ir. J.A. Kors, Dr. D. Rizopoulos en Prof.dr. B. van Calster. Dank voor het kritisch lezen van het proefschrift.

Verder wil ik graag de volgende personen bedanken die tijdens mijn promotietraject met mij hebben meegeleefd, meegeluisterd, meegelezen en meegemopperd:

Allereerst mijn lieve moeder, die het laatste deel van mijn promotietraject helaas niet meer heeft kunnen meemaken. Regelmatig maande zij mij tot spoed omdat zij zo graag bij het einde ervan had willen zijn. Wat had ik dat graag gewild.

Mijn collega's bij mijn diverse werkomgevingen, die altijd belangstelling toonden met betrekking tot de vorderingen van mijn onderzoek.

Mijn lieve vrienden, die vaak tegen wil en dank moesten luisteren naar mijn uiteenzettingen over modelleringstechnieken als support vector machines en random forests en

de waarde van predictiemodellen in een medische context en dat terwijl er ook andere belangrijke zaken aan de orde waren. Ik noem in dezen:

Ron Glandorf, mijn oudste vriend. Wat hebben wij een plezier gehad vanaf het moment dat wij elkaar ontmoetten in 1972 bij de UVA in Amsterdam. Dank voor alle steun en het meeleven.

Nico Visser, mijn oudste vriend qua leeftijd. Als wij wat minder vaak hadden geschaakt, was ik waarschijnlijk veel eerder gepromoveerd en was jouw levenswerk ook allang klaar geweest.

Peter Vaz Nunes en Teun Ammeraal, mijn paranimfen. Dank dat jullie mij terzijde willen staan en dank voor al die mooie en intense momenten van de afgelopen jaren.

## CURRICULUM VITAE

Tjeerd van der Ploeg, MSc, was born in Amsterdam on April 4, 1953. After finishing secondary school at the HBS Van Riemsdijklaan in Beverwijk in 1972, he studied Mathematics and Physics at the University of Amsterdam from 1972 until 1980. Already during his study, in 1975, he started teaching Mathematics at the Berlage scholengemeenschap (secondary school) in Amsterdam, where he worked until 1990. In 1990 he started teaching Mathematics and Statistics at Hogeschool Alkmaar (polytechnic), nowadays Inholland University, where he is still employed today. While working at Hogeschool Alkmaar, he studied Statistics and Econometrics (1992-1995) at the University of Amsterdam. Since 2002, he has been posted by Inholland University as a statistical consultant at the scientific research departments of various peripheral hospitals (Medisch Centrum Alkmaar, Westfries Gasthuis Hoorn, Kennemer Gasthuis Haarlem, Spaarne Ziekenhuis Hoofddorp). During these postings, he supported a great number of medical PhD candidates with the statistical analysis of their research findings. This kindled his interest in medical research and, particularly, in predictive modelling. In 2010, he started a PhD trajectory under supervision of Prof.dr. E. W. Steyerberg at Erasmus MC Rotterdam, department of Public Health, section Medical Decision Sciences. This PhD trajectory was partly facilitated by Inholland University.