Classical Regression and Predictive Modeling

Richard J. Cook[1], Ker-Ai Lee[2], Benjamin W.Y. Lo[3] and R. Loch Macdonald[4]

Affiliations:

1. PhD, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

2. MMath, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

3. MD, Department of Neurosurgery, Lenox Hill Hospital, New York, NY 10075, USA.

4. MD, PhD, Department of Neurological Surgery, University of California San Francisco, Fresno Campus, Fresno, CA 93721, USA.

Corresponding Author:

Richard J. Cook
Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West, Waterloo, ON, N2L 3G1, Canada
E-mail: rjcook@uwaterloo.ca

Keywords: Association, causal analysis, explained variation, classification, prediction, predictive accuracy

Short Title: Classical Regression and Prediction

The final publication of "Cook RJ, Lee K-A, Lo BWY, Macdonald RL (2022). Classical regression and predictive modeling. *World Neurosurgery*, 161, 251-264" is available at Elsevier via https://doi.org/10.1016/j.wneu.2022.02.030 @ 2022. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/.

Abstract

Background: With the advent of personalized and stratified medicine, there has been much discussion about predictive modeling and the role of classical regression in modern medical research. We describe and distinguish the goals in these two frameworks for analysis.

Methods: The assumptions underlying and utility of classical regression are reviewed for continuous and binary outcomes. The tenets of predictive modeling are then discussed and contrasted. Principles are illustrated by simulation and through application of methods to a neurosurgical study.

Results: Classical regression can be used for insights into causal mechanisms if careful thought is given to the role of variables of interest and potential confounders. In predictive modeling, interest lies more in accuracy of predictions and so alternative metrics are used to judge adequacy of models and methods; methods which average predictions over several contending models can improve predictive performance but these do not admit a single risk score.

Conclusions: Both classical regression and predictive modeling have important roles in modern medical research. Understanding the distinction between the two framework for analysis is important to place them in their appropriate context and interpreting findings from published studies appropriately.

1. Introduction

Much scientific research aims to characterize the relationships between individuals' attributes (e.g. demographic or genetic features), their environment (e.g. exposures or treatments), and health outcomes. For this purpose, the attributes and environmental features of interest are summarized in the form of covariates, with the health outcome representing a response in a statistical model. The particular scientific aim may simply be to describe observed relationships between the covariates and the response, or it may be to gain deeper insights into underlying causal mechanisms.[1] Classical regression can be used for both types of objective, but causal analysis typically involves a more formal and exhaustive enumeration of potential confounding variables, careful consideration of the relationships between all variables, and an explicit specification of modeling assumptions; directed acyclic graphs (DAGs) play a key role in organizing and communicating much of this information.[2]

With the advent of personalized medicine over the last two decades,[3] interest has increased in a third type of research aim in medicine - that of predictive modeling.[4,5] In the neurosurgical setting, for example, Byon et al.[6] developed a predictive model for fluid responsiveness in mechanically ventilated children undergoing neurosurgery. Chen et al.[7] considered use of a scoring system developed by Copeland et al.[8] to predict mortality in general neurosurgery that can play a role in audits on performance. In prediction, the goal is not to interpret associations or make causal inferences about the role of key covariates but rather to develop models that can be used to accurately predict health outcomes. Here, relationships between variables are exploited to predict outcomes with a high degree of accuracy, often through derivation of risk scores based on regression models, or more generally from "black box" risk prediction algorithms. In medical settings, decisions are often guided by predictive inferences - individuals designated as "high risk" for poor health outcomes may be offered lifestyle counselling, or more intensive or expensive treatments. Risk predictions also may alter individuals' standing on wait lists for surgical procedures - Kent et al.[9] argued that in settings in which risk prediction models are used to manage health services, they should be transparent, accurate, and updated frequently.

Prediction accuracy is measured differently depending on the nature of the response (i.e. continuous, binary, or censored time-to-event responses). Measures characterizing how close model predictions are to actual outcomes reflect overall performance. A more detailed consideration of predictive performance involves two properties: calibration and discrimination.[10] Calibration refers to the accuracy of estimates of absolute risk[10] in the sense of how well the average predictions align with some marginal population attribute such as the mean response, the prevalence of a condition, or the median time to an event in the failure time setting. Good calibration is key in demand forecasting when predictive modeling issued to plan and budget for health service needs in a population. Discrimination is typically a more important feature of a prediction method in medical settings when focus is on individual patients - it refers to how well a prediction model differentiates between individuals who will experience an event, say, from those who will not experience the event.

The purpose of this article is to review uses of classical regression and explain statistical issues in predictive modeling. We first review modeling for descriptive statistical analyses and causal analyses. We then outline approaches for the development and evaluation of predictive models and prediction more generally. In Section 2, we introduce the linear model and consider the classical aims of regression modeling, which may be descriptive modeling of associations, or the more ambitious and meaningful goal of causal analysis. We then discuss prediction in the context of linear models and ways of measuring predictive performance; this introduces the notions of model fit and explained variation. In Section 3, we consider the setting in which the response of interest is binary and discuss methods for quantifying predictive performance with such outcomes. Remarks on more modern methods for building and evaluating prediction algorithms are given in Section 4. An application is given in Section 5, where the aim is to predict satisfactory long-term outcome among individuals experiencing a ruptured brain aneurysm and undergoing neurosurgery[11]. Issues in predictive modeling with censored time to event data are considered in Section 6, and concluding remarks are provided in Section 7.

2. Classical Regression versus Predictive Modeling

2.1 Descriptive and Causal Analysis via Classical Linear Regression

Classical linear regression involves a continuous response depicted by the letter *Y*, and a set of covariates X_j, $j = 1, \dots, p$. The goal is to study the relation between the covariates and the response through the additive model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \tag{1}$$

where β_0 is called the intercept, β_j is the coefficient of X_j, $j = 1, \dots, p$, and $\epsilon \sim N(0, \sigma^2)$ is a normally distributed error term reflecting random variation. The model is often represented more compactly by defining the covariate vector $X = \left(1, X_1, \dots, X_p\right)'$, the vector of regression coefficients $\beta = \left(\beta_0, \beta_1, \dots, \beta_p\right)'$, and writing $Y = \mu(X;\beta) + \epsilon$ where $\mu(X;\beta) = X'\beta$ is the expected (mean) response for an individual with covariate vector *X*. The model has two parts to it: the systematic component $\mu(X;\beta)$, and the random component $\epsilon = Y - \mu(X;\beta)$, which accommodates variation of the response about its expected value. The standard deviation of the random error, σ, quantifies the extent to which observations can deviate from the expected value, given the covariates; it is often left as an implicit assumption that the error ϵ is independent of the covariates *X*, but as we will see shortly this is an important assumption. With a sample of *n* independent individuals, contributions from each individual represent independent realizations of the joint process generating *(Y, X)*. If the relationship between the covariates and the response is of interest, we interpret β_j as the expected change in the mean response when we increase X_j by one unit, when all other covariates are held fixed; often we use the term “when controlling for all other covariates”. If interest lies in assessing whether X_j adds any explanatory power in the presence of the other covariates, a test of the important of X_j can be carried out by testing whether or not $\beta_j = 0$ in (1).

To simplify the discussion, suppose *p = 2* giving

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$. If (2) is viewed as the "correct model" then β_1 is the causal effect of a one-unit increase in X_1 on the expected response when X_2 is held fixed. If X_2 is omitted from the model and

$$Y = \alpha_0 + \alpha_1 X_1 + \epsilon; \tag{3}$$

is fitted, one can show that

$$\alpha_1 = \beta_1 + \beta_2 \Delta_2, \tag{4}$$

where $\Delta_2 = E(X_2 | X_1 = 1) - E(X_2 | X_1 = 0)$ and $E(X_2 | X_1 = x_1)$ represents the average value of X_2 when X_1 takes on a particular value x_1. The term $\beta_2 \Delta_2$ represents the bias of the estimator of α_1 in (3) for the causal effect β_1 in (2). This bias arises from the confounding effect of omitting X_2 when it is associated with the response and the covariate X_1. Of course, if X_2 is not important (i.e. $\beta_2 = 0$) or if X_1 and X_2 are not associated, the bias is zero; in the latter case $\Delta_2 = 0$, since the average value of X_2 will be the same regardless of the value of X_1 (i.e. $E(X_2 | X_1 = 1) = E(X_2 | X_1 = 0)$). More generally, however, the inference we draw regarding the importance of X_1 when we fit (3) is influenced not just by its true causal effect β_1, but also by the effect of X_2 in (2) denoted by β_1, and the strength of the association between X_1 and X_2. The complex nature of this bias makes the estimator for α_1 difficult to interpret meaningfully which has lead to the greater appreciation of the importance of causal thinking. Thus, when analyses are carried out more informally, they are best viewed as descriptive, and inferences should be confined to statements about associations between variables. Of course, the underlying model (1) adopted here is very simple, and in practice there are many more potential confounding variables at play. The use of naive simply models in more complex settings yield estimators that are subject to influence from many confounders, making interpretation even more difficult than illustrated here.

Returning to the simple setting of (1), we note that in a randomized clinical trial, we may have $X_1 = 1$ or 0 if an individual is assigned to the experimental intervention or standard care, respectively. In this setting, X_2 may represent an important prognostic variable. Because X_1 is assigned by randomization, it is independent of X_2, and so simply fitting the model (3) will yield an unbiased estimate of the causal effect β_1 and a simple t-test can be performed. A test for the treatment effect based on (2) corresponds to an analysis of covariance where adjusting for X_2 will often reduce the residual variation and hence can increase the power of the test of treatment effect.[12]

2.2 Predictive Inference in the Linear Model

In prediction, interest lies less in studying the relationship between particular covariates and the response. Instead, the goal is to use available covariate information to obtain a predictive estimate of a response, sometimes by a simple model incorporating available contextual information, and sometimes by complex and non-transparent black-box computing algorithms. Here, we revisit the regression model (1) and note that the mean $\mu(X;\beta)$ is the expected value of the response for an individual with a given set of covariates X; the arguments X and β make it clear that the mean depends on both the covariates and their corresponding regression coefficients. Upon fitting a regression model, we have an estimate $\hat{\beta}$ and we estimate the mean as $\mu\left(X;\hat{\beta}\right) = X'\hat{\beta}$. This estimate, being the best estimate of the expected response, is used for prediction. Thus, if a new individual is encountered with covariate X_{new}, their predicted response would be $\hat{\mu}_{new} = X'_{new}\hat{\beta}$ and the error between their actual response and the predicted response would be $\epsilon_{new} = Y_{new} - \hat{\mu}_{new}$. Since the goal of predictive modeling is to use covariates to posit a value that will be close to the response as possible, we want to select predictive models that minimize the size of ϵ_{new} across individuals in a target population.

Schemper[13] introduces the general measure *D* to quantify the variation of the random component of a simple model involving no covariates, where *D(X)* is the corresponding measure in a regression model including covariates $X_1, \dots, X_p$. Then, *[D -D(X)] = D* is the proportion of the variation in the null model that can be explained by the covariates in the model using *X*; this quantity falls between zero and one. If *p* is large and knowledge was sufficient that all relevant covariates were available, then the percentage of variation would be very high, but typically there remains a large amount of unexplained variation and the proportion of variation explained will be modest. The predictive accuracy of a model with a very high proportion of explained variation will be excellent, and covariate information from a fitted model will be very helpful in anticipating a response. The sum of squared errors (or the error sum of squares) is a common measure used for *D*. In this case, the total sum of squares (SS_{tot}) is for a model with no covariates, and the residual sum of squares (SS_{res}) is the corresponding terms for the predictive model involving covariates. The proportion of variation explained is then $PVE = (SS_{tot} - SS_{res}) = SS_{tot}$; this is sometimes labeled R^2, as we do in what follows and is called the coefficient of determination. If the covariates explain the variation in the response very well predictions will tend to be close to the responses and *PVE* will be large. Note that a covariate may have a large and significant regression coefficient but paradoxically fail to explain much variation in the response and therefore not enhance predictive accuracy. This arises when there is little variation in the covariate in the sample - some variation is needed to detect a significant effect of course, but it may be of limited value for prediction if the vast majority of individuals share the same value of the covariate.

To illustrate, consider a simple linear regression model with

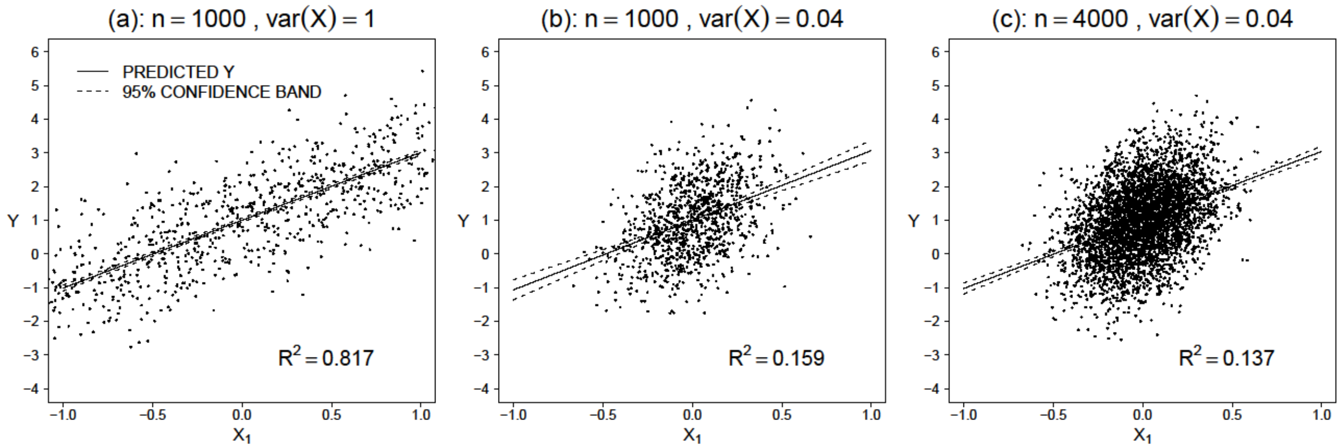
$$Y = \beta_0 + \beta_1 X_1 + \epsilon, \tag{5}$$

with $\epsilon \sim N(0,\sigma^2)$, $\beta_0 = 1$, $\beta_1 = 2$, and $\sigma = 1.0$. We consider Scenario A, in which X_1 is normally distributed with mean zero and $var(X_1) = 1.0$. In this linear model, the *PVE* is given $\beta_1^2\text{var}(X_1)/(\beta_1^2\text{var}(X_1) + \sigma^2)$, which here gives *PVE* = 0.80. We simulate *n = 1000* observations $\{(Y_i, X_i), i = 1, \dots, 1000\}$ and create a scatter plot of *Y* versus X_1 in Figure 1(a). Scenario B is identical, except we greatly reduce the variation in the covariate and set *var(X_1) = 0.04*, giving a *PVE* = 0.138; this substantial drop from Scenario A is due to the smaller variation in X_1. The scatterplot of a sample of *n = 1000* observations is given in Figure 1(b), where the empirical estimate of the *PVE* is given as 0.159; note that the scales for Figures 1(a) and 1(b) are the same to make the contrasts clear. The first two columns of Table 1 report the estimated regression coefficients for these two simulated datasets, and The fitted regression lines and 95% confidence intervals (CIs) are superimposed on the corresponding scatterplots. The estimated regression coefficients for X_1 are 2.013 and 2.067 for Figures 1(a) and 1(b), respectively (close to the true value of 2). Also, note that the standard error of the regression coefficient for X_1 is 0.030 for Scenario A and much greater at 0.151 for Scenario B; this reflects the impact of the variation in the covariate on the precision of the estimated coefficients - the greater the variation in the covariate, the more precise the estimated regression coefficient.

Table 1: Table of estimated regression coefficients, standard errors, and p-values for Scenarios A, B and C

	Scenario A *N = 1000, var(X_1) = 1*			Scenario B *N = 1000, var(X_1) = 0.04*			Scenario C *N = 4000, var(X_1) = 0.04*		
	Est	SE	p	Est	SE	p	Est	SE	p
β_0	1.004	0.031		1.004	0.031		1.012	0.016	
β_1	2.013	0.030	<0.001	2.067	0.151	<0.001	2.034	0.081	<0.001

Figure 1: Scatter plots and fitted regression lines for Scenarios A, B and C, where the proportion of variation explained is 80%, 13.8% and 13.8%, respectively.



In Scenario C, we retain the parameter settings of Scenario B, but quadruple the sample size to *n = 4000*; the corresponding scatterplot is given in Figure 1(c) and the estimates are reported in the last column of Table 1. Here, we see that the point estimate for β_1 is again close to the true value of 2, but the standard error is about half of what it was in Scenario B. This reflects the general phenomenon that the standard error decreases linearly with $1/\sqrt{n}$ where *n* is the sample size - thus by quadrupling the sample size we halved the standard error. This also highlights the fact that the power to detect an important effect is influenced by both the sample size as well as the variation in the covariate; this is well known in experimental design but is seldom discussed explicitly in the analysis of observational data. The proportion of variation explained is comparable in Figures 1(b) and 1(c) at 0.159 and 0.137, respectively, also highlighting the fact that measures of predictive performance are (as they should be) independent of sample size. To summarize, in each of Scenarios A to C the regression coefficients are significant, but in Scenarios B and C the proportion of explained variation is low because there is little variation in the covariate.

Other measures of predictive accuracy that reflect discriminatory power of a predictive model include Harrell's C index.[14] This method considers all possible pairs of individuals in a data set and computes the proportion of such pairs for which the ordering of the predictions is concordant with the ordering of the responses - high values (close to 1) represent good discriminatory power, whereas values around 0.5 indicate poor discriminatory power.

3. Prediction with Binary Outcomes

3.1 General Framework for Modeling

In many settings, the response of interest is binary, reflecting whether a surgical intervention was successful or not, or an event or complication was experienced during follow-up. If Y denotes a binary variable taking the value 1 if an outcome occurred and 0 otherwise, generalized linear models[15] may be formed based on the probability $P(Y = 1 \mid X)$. For binary responses $P(Y = 1|X) = E(Y|X) = \mu(X)$ is the mean given X and upon specification of a link function $g(.)$ we set

$$g\big(\mu(X;\beta)\big) = \beta_0 + X'\beta, \tag{6}$$

where X is a $p \times 1$ covariate vector and β is a $p \times 1$ vector of regression coefficients.[16] Logistic regression models are perhaps most often used where $g(\mu) = logit(\mu) = log(\mu/(1-\mu))$. Upon fitting this model and obtaining estimates $\hat{\beta}$ we can think of $\hat{\beta}_o + X'\beta$ or $\hat{\mu}(X) = \mu\big(X;\, \hat{\beta}\big) = g^{-1}\big(\hat{\beta}_0 + X'\hat{\beta}\big)$ as a risk score. Since the latter is based on an estimated probability it is called a *probabilistic prediction*, which is simply the estimated probability that the response will be 1. With a threshold *c* specified, the corresponding *point prediction* would be $\hat{Y} = I(\hat{\mu}(X) > c)$where $I(.)$ is an indicator function such that $I(A) = 1$ if A is true and is zero otherwise. The key distinction is that a probabilistic prediction will not take on values that are possible for the actual response whereas point predictions will.

As in the case of continuous responses, with binary responses measures of predictive accuracy reflect how close predictions are to realized responses which is referred as overall performance. One common measure proposed by McFadden[17] is

$$R_{MF}^2 = \frac{\sum_i l_i(\mu_0) - \sum_i l_i(\mu(x_i;\hat{\beta}))}{\sum_i l_i(\mu_0)}$$

where μ_0 is the proportion of individuals with *Y = 1* and $l_i(\mu) = y_i \log \mu + (1 - y_i)\log(1 - \mu)$. Several variations of this formula have been proposed, including one by Nagelkerke[18] that involves a standardization to ensure an upper bound of 1 can be reached; see Menard[19] for further discussion of this family of measures of predictive performance.

An alternative is to consider a measure of predictive accuracy based on the absolute performance. The Brier score is a measure of predictive accuracy for categorical outcomes and probabilistic prediction.[20] If $y_1, \dots, y_m$ denote observations from a validation sample and a prediction model obtained from a *training sample* yields probabilistic predictions $\hat{p}(X_i)$, then the Brier score (B) is defined as

$$B = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \mu\left(x_i;\hat{\beta}\right)\right)^2. \qquad (7)$$

In this case, a predictive model is perfect if the Brier score is zero. The upper bound of a scaled Brier score can be set to 1, but the precise interpretation of the Brier score (or its scaled version) is unclear; Brier scores are useful for comparing predictive models (lower values correspond to better predictive models), which can be done informally or formally.[21]

With binary outcomes discrimination refers to how well a predictive model differentiates those at higher risk of an event from those at lower risk, whereas calibration refers to how well the predictive model reproduces marginal features of a population distribution.[10] A predictive model that yields predictions that, when averaged, reflect the population mean response is well-calibrated. If it does not yield predictions that are close to individual responses (i.e. if it does not discriminate well between people with and without an event) it is not useful for decision-making at the individual level. The concept of overall performance of a predictive model is based on how close predictions are to responses and usually assessed by Brier scores using probabilistic predictions.[22] The Hosmer-Lemeshow test[23] is often used for assessing fit of a model to the dataset used to build it. We expand our discussion of discrimination-based measures of predictive accuracy in the following section.

3.2 Measures of Discrimination

Point Prediction versus Probabilistic Prediction

Note that for continuous outcomes the range of possible values for the response and the prediction are compatible. For binary or other discrete responses, risk scores may take on values that are not possible for the response. Point prediction is the term used to describe a prediction taking on the possible values of the response. With a binary response a natural point prediction $\hat{Y}$ based on the probabilistic predictor $\hat{\mu}(X) = P(Y = 1|X;\hat{\beta})$ is $\hat{Y} = I(\hat{\mu}(X) \geq 0.5)$; that is, for a point predictor we would predict the value that has the greatest probability of occurrence based on our model, but thresholds other than 0.5 are often preferable as we next describe.

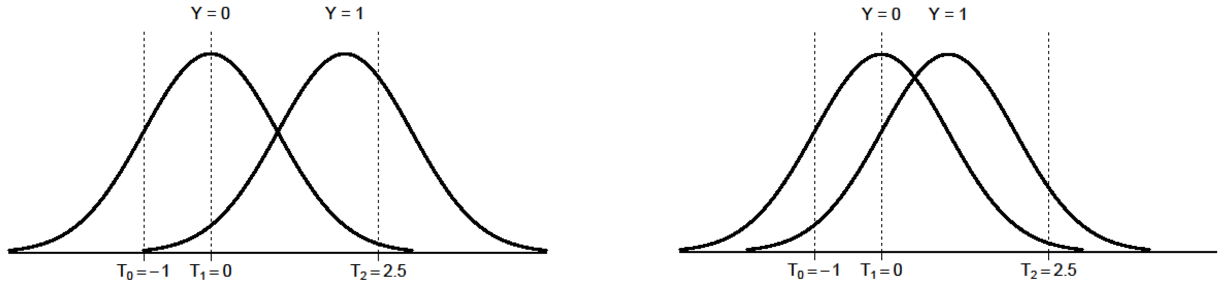
Table 2: Cross-classification of individuals according to their point predictions and responses in the form of a confusion matrix with threshold T

	Y = 1	Y = 0	Total
$\hat{Y} = 1$	A(T)	B(T)	A(T) + B(T)
$\hat{Y} = 0$	C(T)	D(T)	C(T) + D(T)
Total	A(T) + C(T)	B(T) + D(T)	

Having developed a risk score one can assess how well the risk score performs in predicting outcomes. This is typically done by considering a range of possible thresholds for the risk score, assigning point predictors accordingly, and examining the concordance between the point predictor and responses. So, if we use the probability scale for the risk score, we might choose a threshold T (0 ≤ T≤ 1) such that if $\hat{\mu}(X) > T$ we predict $\hat{Y}(X) = 1$ and we set $\hat{Y}(X) = 0$ otherwise. We can then create a 2×2 table (Table 2) referred to as a "confusion matrix" in machine learning,[24] which shows the relation between the true responses (*Y*) and predictions ($\hat{Y}$). With a given threshold *T*, *A(T)* is the total number of individuals for whom their response is 1 and their point predictor is 1 (i.e. their point predictor and response are concordant at 1), while *B(T)* is the total number of individuals for whom their response is 0 but their point predictor is 1 with the threshold *T*, representing a misclassification; the other cells are similarly defined.

From this we calculate estimates of the *false positive rate* (FPR) of the point prediction with threshold T, defined as the proportion of individuals for whom *Y = 0* but who were classified with $\hat{Y} = 1$, given here by *FPR(T) = B(T)=(B(T)+D(T))* and the *false negative rate* (FNR) estimated as *FNR(T) = C(T)=(A(T) + C(T))*. The sensitivity of the classification scheme is the complement of the FNR, sometimes called the true positive rate (TPR), and is estimated as *SENS(T) = A(T)=(A(T) + C(T))* and the specificity is the complement of the FPR, and is given by *SPEC(T) = D(T)=(B(T) + D(T))*. When values of *T* range from 0 to 1 all possible thresholds are considered and one can plot the TPR against the FPR to form *receiver operating characteristic* (ROC) *curve*; see Pepe[25]. A classification rule that is of no predictive value will generate a 45-degree line on the ROC curve, whereas risk scores yielding better discrimination will generate curves above the 45-degree line; the better the discrimination, the greater the curve, motiving the use of the area under the ROC curve (AUC) as a summary statistic of discriminative ability of the risk score - this is sometimes referred to as the C-statistic.[26] The C-statistic represents the probability of concordance, that is, the probability that randomly chosen pair of subjects, one with *Y = 1* and one with *Y = 0*, are both correctly classified yielding a concordance of their true and predicted classes. Note, however, that the ROC curve is constructed by considering all possible thresholds, but a particular threshold will often need to be chosen when making predictions; the ROC is therefore a reflection of the potential discriminative ability of predictions based on the risk score.

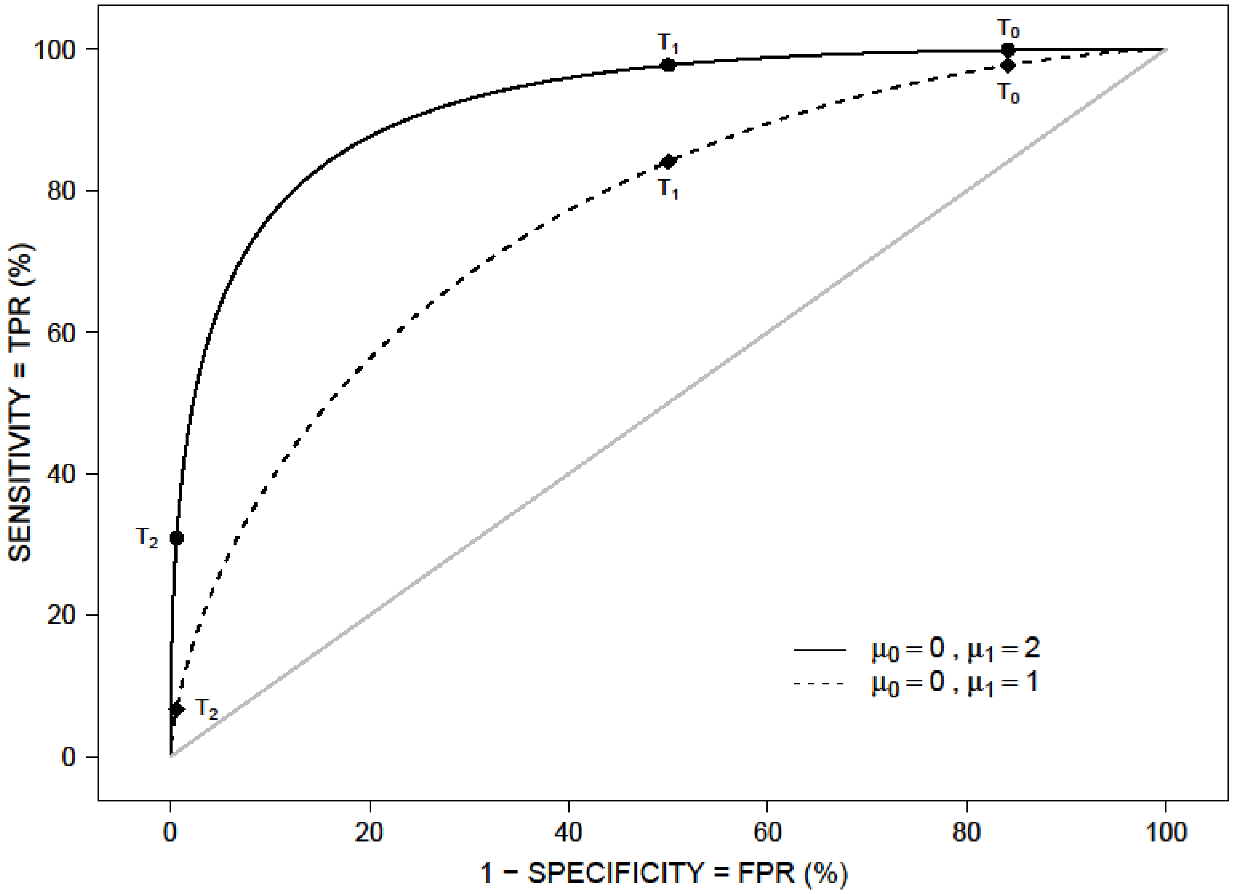
Figure 2: Risk score distributions for a setting with good (panel (a)) and poorer (panel (b)) discrimination; Risk score is standard normal N(0; 1) for Y = 0 in both scenarios



(a) Risk score distribution is $N(\mu_1=2, \sigma= 1)$ for *Y = 1*

(b) Risk score distribution is $N(\mu_1=1, \sigma= 1)$ for *Y = 1*

Figure 3: Receiver operating characteristic curve for the setting of Figure 2(a) (higher curve) and Figure 2(b) (lower curve)



Consider the distribution of risk scores for individuals with *Y = 0* and *Y = 1*, where higher values of the risk score tend to occur for those with *$Y = 1$*. For illustration, we suppose that the risk score is normal with mean 0 if *$Y = 0$* and mean 2 if *$Y = 1$*, with the standard deviations common at 1; the distributions are plotted in Figure 2(a) where we overlay three illustrative thresholds for classification at *$T_0 = -1$*, *$T_1 = 0$* and *$T_2 = 2.5$*. With threshold T_0, the probability an individual with Y = 0 has a risk score above -1 is denoted by *A_0*, so we write *$FPR(-1) = FPR_0 = A_0$*, and the probability an individual with *$Y = 1$* has a risk score above -1 is *B_0*, so we write *$TPR(-1) = TPR_0 = B_0$*. With threshold *T_1* the probability an individual with *$Y = 0$* has a risk score above 0 is A_1 so we write FPR(0) = *$FPR_1 = A_1$*; the probability an individual with *$Y = 1$* has a risk score above 0 is *B_1*, so we write *$TPR(0) = TPR_1 = B_1$*. With threshold *$T_2 = 2.5$*, the probability an individual with *$Y = 0$* has a risk score above 2.5 is *A_2*, giving *$FPR(2.5) = FPR_2 = A_2$*, and the probability an individual with Y = 1 has a risk score above 2.5 is *B_2*, so we write *$TPR(2.5) = TPR_2 = B_2$*. These points are plotted in Figure 3 with the labels *T_0*, *T_1* and *T_2* to indicate the threshold to which they correspond. All such thresholds can be considered, and if we plot the TPR (%) against the FPR (%) for each possible threshold, we construct a ROC curve depicted by the solid curve in Figure 3. The greater the separation between the two risk score distributions, the better the risk score serves as a basis for classification. Figure 2(b) displays risk score distributions when there is less separation, which gives the lower ROC curve in Figure 3 with the dashed line. The discriminatory ability, which is the focus of binary classification model, is typically characterized by sensitivity, specificity, and the concordance statistic (C-statistic), which is equivalent to the area under the ROC curve.[27]

3.3 Over-fitting and Issues in Validating Models

Prediction models are obtained by modeling the relationship between covariates and a response in sample data. Such modeling procedures typically aim to find the best model among a family of possible models. Traditional methods involving forward or backward selection yield biased estimates of the effects of selected covariates due to the “winner's curse”.[28] This refers to the phenomenon by which over-fitting lead to the inclusion of covariates that enhance fit for the particular sample but results in the model not performing well in other samples, even if randomly drawn from the same target population. Zhong and Prentice[29] describe methods for bias-reduction that are important when a large number of variables are being considered. Ridge regression was proposed as a means of smoothing regression coefficients to attenuate larger estimates due to unique features of a particular random sample.[30] We comment on penalized regression and other more modern machine learning methods in Section 4.

When a single sample is available for building and assessing a predictive model, more honest assessments of prediction accuracy can be obtained by dividing a sample into a training sub-sample (say 50% - 75% of the original sample) used to build a predictive model, and the complementary sub-sample to be used to assess predictive accuracy - this technique is called *split-sample validation*. This approach is criticized as poor because the model is built on a smaller sub-sample that may have unique features by chance from the random splitting process, and that the validation sample is correspondingly modest. It is particularly problematic with binary outcomes when the overall event rate is low. Some of the limitations of this approach are addressed by use of *K*-fold cross-validation wherein the sample is dividing into *K* distinct sub-samples (or “folds”) of equal size, the model is built in each of K training subsamples, each comprised of the distinct union of *K- 1* folds, and the predictive accuracy is assessed in the fold that was held out from the training sample. For this approach, the model building benefits from the larger training sub-samples, and the full sample is ultimately used to assess prediction accuracy. Note that with *K*-fold cross-validation, we end up with *K* different predictive models, so this really is assessing the model building procedure in the context of the available data, rather than a single predictive model that would arise from the split-sample validation. It therefore has a useful role in selecting tuning parameters for penalization methods where the question is a more global question about how to avoid over-fitting, rather than assessing performance of a particular model. An extreme form of *K*-fold cross-validation is when $K = n$, where n is the sample size - this is called the *jack-knife approach*.

Steyerberg[5] points out the bootstrap procedure is preferred method as the resampling recognizes the fact that the sample was drawn from a population. Here, we may take *B* bootstrap samples of size *n* by resampling with replacement, build the predictive model for each bootstrap sample, and assess the predictive performance in the bootstrap sample on which it was built that estimates the apparent predictive accuracy, and in the original sample. The difference of the two measures of performance is a measure of the optimism of predictive performance when assessing validation in the same sample as the sample used to build the model. It can therefore be used to “correct” for the optimism one might otherwise have about performance.

The critical role of the covariate distribution in predictive performance of a model was mentioned in Section 2.2. When models are built for prediction in a given setting and to be used in a population with a quite different covariate distribution, performance will often be compromised - this feature of a predictive model is called “transportability”. As an example, a model built in a sample with an equal proportion of men and woman may find sex to be a variable in a risk score. If the same model is to be used in a sample comprised exclusively of men, then this element of the risk score will not be informative and so some degradation of performance can be expected. Use of external validation samples can help give insight into the generalizability of predictive performance metrics. Careful consideration of the composition of the external validation sample will be necessary to determine if the findings from this validation exercise are influenced by this factor. If interest lies in estimating the predictive performance of a model in a specific setting for which the covariate distribution can be specified, adjusted measures of predictive performance can be calculated.[31]

4. Penalized Regression and Machine Learning Techniques

4.1 Penalized Regression

Breiman[32] noted that the traditional methods of best subset selection (e.g. forward or backwards elimination) yielded unstable models and that such instability could lead to poor predictive performance. An alternative approach to model building is to define an objective function to be maximized but to penalize this function to mitigate over-fitting. Ridge regression[33] imposes some shrinkage, which leads to more stable models, but does not set any coefficients to zero and therefore does not “select” key variables. The LASSO[34] is a penalization approach wherein the penalty attempts to maintain the advantages of both subset selection and ridge regression by shrinking some coefficients and setting others to zero through use of a log-likelihood with a so-called L_1 penalty function. This has the form

$$l(\beta) - \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \qquad (8)$$

where $l(\beta)$ is the log-likelihood and the second term is the penalty. This model is selected by maximizing this function but each time a coefficient deviates from zero the function decreases by an amount $\lambda|\hat{\beta}_j|$, which combats over-fitting. Other penalty functions which have recently been proposed include the smoothly-clipped absolute deviation (SCAD)[35,36], the adaptive LASSO[37], the elastic net[38], the grouped LASSO[39], and the minimax concave penalty (MCP)[40]. While much of the work on variable selection techniques was initially carried out in the context of continuous responses, advances have been made to deal with binary responses and time-to-event responses. For the latter, the penalty term is typically applied to the partial likelihood arising from a semiparametric Cox regression model[41] when data are right-censored.

4.2 Classification and Regression Trees

An alternative nonparametric framework for the development of prediction algorithms is through recursive partitioning. In this framework, a sample is initially sub-divided to create two distinct sub-samples of individuals in which individuals in different sub-samples are dissimilar with respect to a chosen measure, that is, individuals within the same sub-sample are more similar to one another than individuals in the opposing partition. Given a list of covariates, the algorithm considers splits based on each possible categorization using all discrete and continuous covariates where for the latter type of covariate every possible cut-point is considered. The optimal split is the one that leads to the greatest separation between the sub-samples (i.e. the two branches of a tree). Often, this decision is based on the minimum p-value or the maximum test statistics[42] among the candidate tests but other criteria can be applied[43]. Following an initial binary split, the two sub-samples are considered for further partition by splitting on another variable within each sub-sample leading to the term recursive partitioning. The sequence of divisions is repeated, with the resulting splits depicted graphically in a tree formation; this stage is called the tree-growing step. Cross-validation is typically used to gauge the extent of over-fitting during which results from excessive splitting. The concern here is that the partitioning is carried out on a single dataset to make use of the information in that particular sample – if this is done to an excessive degree, then it may perform well in addressing the idiosyncrasies of that sample at hand, but not other samples which may be drawn from the same population but differ in some random way. The tree pruning stage uses cross-validation to produce a final tree that balances simplicity and fit, and in so doing renders a classification procedure that may be more generalizable and perform better in external samples.

Recursive partitioning may be viewed as a nonparametric procedure since there are no regression coefficients naturally associated with the final tree and splits may be based simply on test statistics. Instead of obtaining regression coefficients we obtain samples of individuals in the terminal nodes (a term for the sub-samples of individuals where no further partitioning is carried out) are viewed as similar with a common prediction for them. If a new individual is encountered, the prediction algorithm based on recursive partitioning simply determines which terminal node they fall in, and assigns the predictive value as the average value among individuals in the respective terminal node. For binary outcomes, this corresponds to the proportion of individuals who have the response of interest, which is an estimated probabilistic predictor - based on this a point predictor can be obtained as described earlier. For binary responses, the term used is a classification tree, but both classification and regression trees are subsumed in the acronym CART. See Strobl et al.[44] for an accessible review.

4.3 Ensemble Methods

Ensemble methods recognize that individual predictive models may suffer from instability in the selection and estimation of covariate effects, thus better performance may be obtained by considering several predictive models and averaging prediction over the set of predictive models. A particularly popular ensemble algorithm is based on the growth of several classification trees to form a collection of trees - this algorithm is called *random forests*. A prediction is then made from each tree and the predictions are aggregated to obtain an overall prediction. The idea of adopting multiple prediction models is to avoid over reliance on a single classification scheme when other classification or prediction rules may perform quite closely in terms of overall performance.

The algorithm of growing a tree will produce the same tree unless the datasets are different. To address this training data are resampled with replacement via the nonparametric bootstrap to create B bootstrap samples of the same size as the original. For boosting methods, the recursive partitioning algorithm is applied to each bootstrap sample and the prediction is made by averaging the predictions for each of the B trees grown. Random forests work in a similar way, but involve another random element - the set of covariates considered for growing the tree for each bootstrap sample is a random subset of the full set of covariates - this ensures that there is more diversity in the types of trees that are grown across the bootstrap samples.

5. Prediction Following Surgery for Ruptured Aneurysm

Here we consider illustrating the selection and evaluation of predictive models and algorithms for a binary outcome based on the Tirilazad database.[11] This data arose from multicenter randomized double-blinded and placebo-controlled trials conducted between 1991 and 1997 involving adult patients with evidence of ruptured brain aneurysms and associated subarachnoid hemorrhage (SAH). The outcome of interest is the dichotomized patient's Glasgow Outcome Score (GOS) 3 months' post-rupture; the response is an indicator of good recovery with potentially moderate disability but functional independence, whereas a poor outcome includes the events of death or loss to follow-up.[11] Thirty covariates displayed in Table 3 are considered for the development of predictive methods. For the analyses that follows, we make use of data from 3,137 patients with complete covariate information and split the data randomly into a training sample (75%) and a type of external validation sample (25%) used to assess predictive performance. All analyses were implemented using the R language (Version 4.1.1).[45]

Table 3: A list of 30 covariates considered for the prediction of subarachnoid hemorrhage (SAH) from Lo et al.[11]

Demographic variables:	**Treatment related neurological variables:**
Age (years)	Time to surgical treatment (hours)
Sex	Treatment arm receiving Tirilazad
Weight (kg)	Severe vasospasm needing balloon angioplasty
Non-treatment related neurological variables:	**Systematic variables:**
Hospital admission neurological (Hunt and Hess) grade	Systolic blood pressure (mmHg)
Aneurysm size ≤ 12 mm	Diastolic blood pressure (mmHg)
Presence of admission angiographic vasospasm	Occurrence of fever one week after admission
Presence of intraventricular hemorrhage	History of hypertension
Presence of intracerebral hematoma	History of angina
Posterior circulation location of aneurysm	History of myocardial infarction
Subarachnoid hemorrhage thickness ≤ 1 mm	History of diabetes mellitus
Prior episode of SAH	History of hepatic disease
History of migraines	History of thyroid disease
Presence of hydrocephalus	Development of pulmonary edema
Development of cerebral edema	
Occurrence of post admission stroke	
Development of vasospasm during treatment	
Seizures requiring antiepileptic medications	

We consider the LASSO via the glmnet package[46] wherein the penalty parameter λ in (8) is set using 10-fold cross-validation via the cv.glmnet function with a view to minimizing the error based on loss functions involving the AUC, deviance or misclassification error. For the CART analysis we used the rpart package[47], where we specify the minimum number of observations in each terminal node as 10, and set the complexity parameter to 0; 10-fold cross-validation was again used (Figure 4). The pROC package[48] was used to create the ROC curve corresponding to the final tree. The random forests algorithm was implemented using the randomForest package[49] with 100 trees grown. To assess the importance of different covariates we consider the mean decrease of accuracy (i.e. how much accuracy the model losses by excluding each variable) and the mean decrease of the Gini index (i.e. this indicates how each variable contributes to the homogeneity of the nodes at the end of the tree)[43] - the more important the variable is for the respective criteria, the greater it is in the plot; see Figure 5. It is apparent that the most important variable is the history of myocardial infarction, followed by neurological grade at hospital admission.

Figure 4: The final CART following pruning according to the optimal complexity parameter

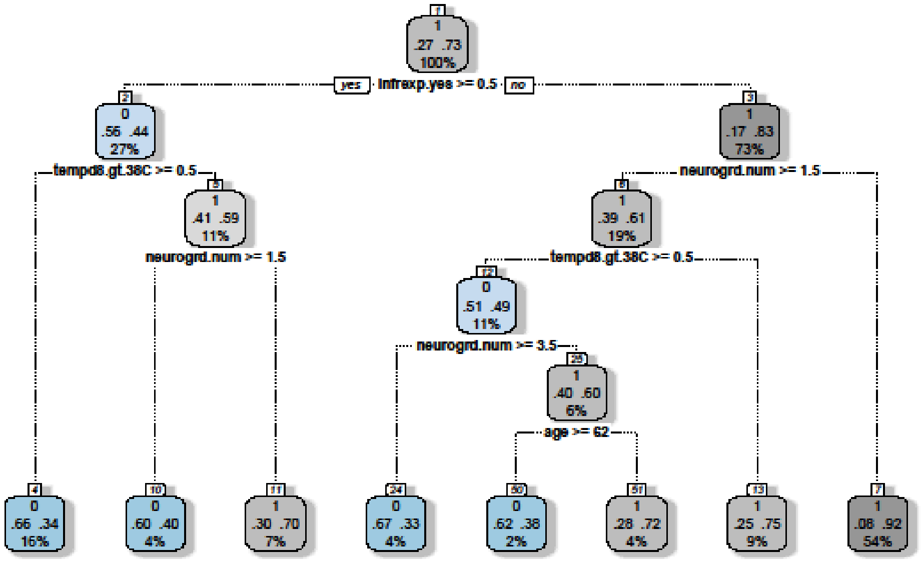
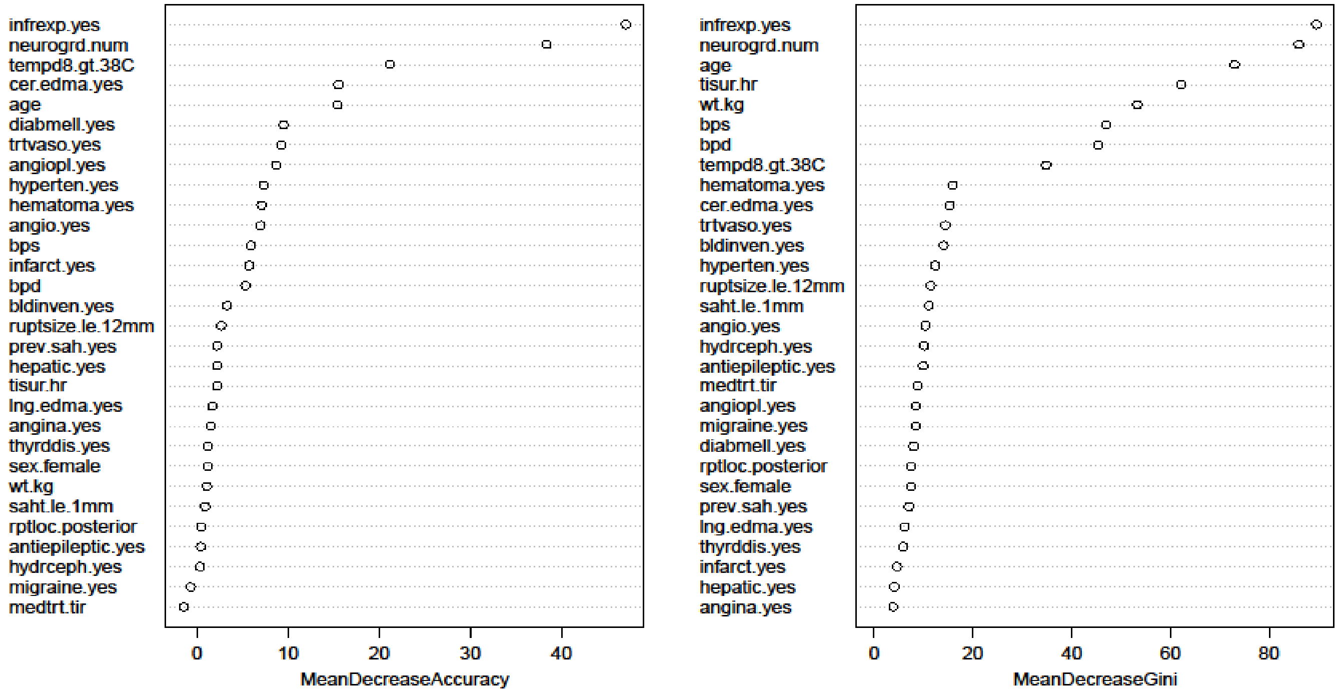


Figure 5: Plots reflecting importance of covariates based on 100 random trees: The *left panel* is based on mean decrease in accuracy, and the *right panel* is based on mean decrease in node impurity measured by the Gini index.



The ROC curves corresponding to each of the predictive analyses are given in Figure 6, where it can be seen that the LASSO performs well with an AUC of 85.6% (95% CI: 82.9, 88.3). The CART algorithm does a poorer job in this study, with an AUC of 80.4% (95% CI: 77.1, 83.6), but the random forest algorithm yields superior discriminatory performance with an AUC remarkably close to that of the LASSO with a value of 85.5% (95% CI: 82.7, 88.3); See Table 4.

Table 4: Predictive performance of LASSO, CART, and random forest approaches on the validation data from Lo et al.[11]

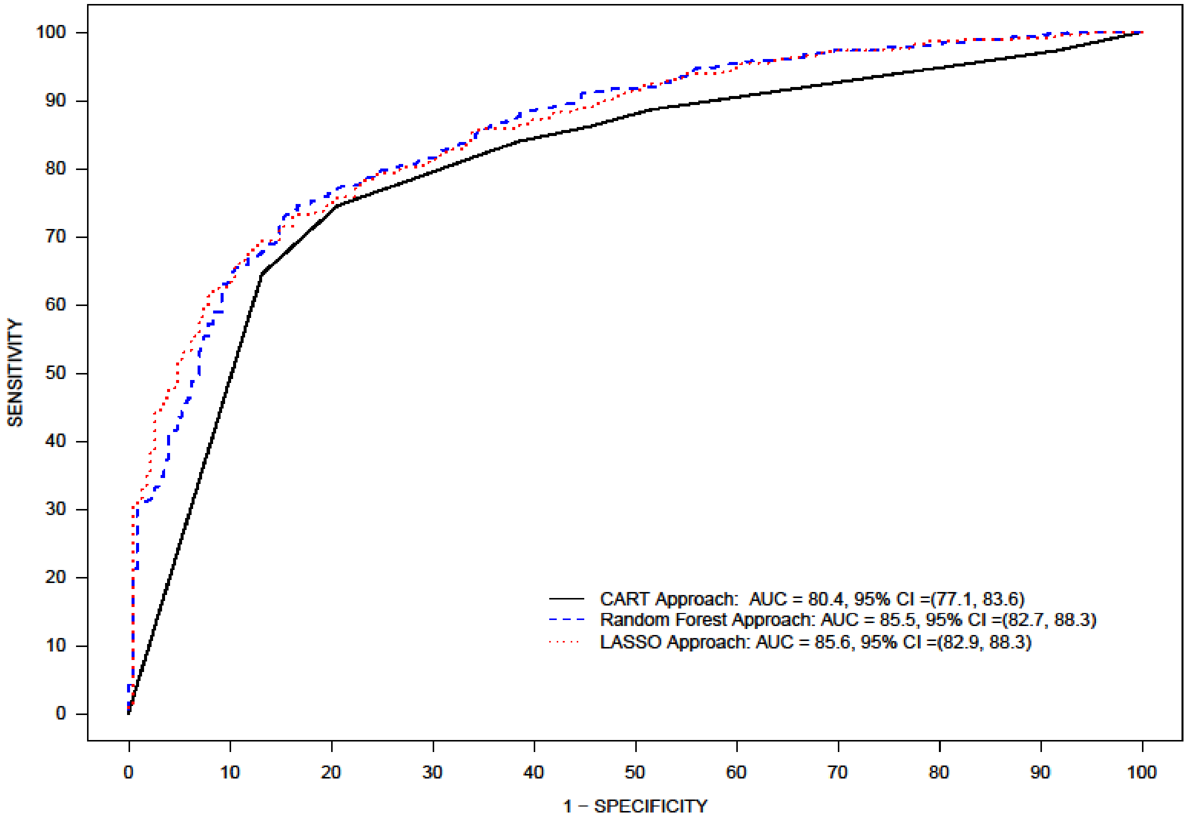
Approach	AUC (95% CI)	Accuracy*	Classification Error Rate†	Brier Score‡
LASSO	0.856 (0.829, 0.883)	0.794	0.206	0.137
CART	0.803 (0.771, 0.836)	0.774	0.225	0.153
Random Forest (100 trees)	0.855 (0.827, 0.883)	0.802	0.197	0.139

* Accuracy = (A(T) + D(T)) = (A(T) + B(T) + C(T) + D(T)); See Table 2.

† Classification error rate = 1 - Accuracy.

‡ Brier score can be calculated using the BrierScore function in the DescTools package[50].

Figure 6: Receiver operating characteristic (ROC) curves and associated areas under the curves (AUCs) corresponding to use of the LASSO, CART, and random forest.



6. Some Remarks on Prediction with Time to Event Responses

When interest lies in longer-term outcomes, prediction methods for time-to-event responses must accommodate censoring due to loss to follow-up both in the training and validation samples. For training samples, a natural starting point is to fit a Cox regression model or build one using penalized regression; the LASSO is commonly used with Cox regression[51] and was among the first to be developed, but many of the common penalty functions of interest can be applied in this setting. Classification and regression trees also have played a role in risk prediction, with early papers based on Cox regression or more nonparametric approaches to tree growing based on the log-rank statistic, for example[52], see also LeBlanc and Crowley[53] and Molinaro et al.[54] and Steingrimsson et al.[55] for more recent developments. Early work on ensemble methods for variable selection and prediction based on the Cox model includes Zhu and Fan[56]; Ishwaran et al.[57] introduces new splitting rules and describes innovations for dealing with incomplete data.

Prediction models may be directed at predicting event times or, more commonly, the event status of individuals at a specified time horizon t_o. In either case, assessing predictive performance with censored data can be challenging, since times of interest may be unknown in the validation sample due to right-censoring; likewise, when the goal is to predict event status at t_o, this will be unknown for individuals who were event-free censored at some time $A < t_o$ where A represent an administrative censoring time. Methods for dealing with unknown event status at a particular time horizon of interest include imputation-based techniques, or use of inverse probability of censoring weights which address the use of a biased sub-sample when attention is restricted to individuals whose data are complete enough to use them in the validation exercise. This requires modeling the censoring time distribution given fixed or possibly time-varying covariates. Harrell's C index[58] is another measure of predictive discrimination where in the failure time setting one again assesses concordance in the ordering between the observed and predicted times within each pair of individuals in a data set. That is the failure time and predicted failure time are later for one member of a pair than the other member, the pair would be designated as exhibiting concordance, but not otherwise. The proportion of all pairs for which the predicted and realized orderings are concordant defines Harrell's C index. The challenge with censored data is that individuals' failure times may be unobserved due to censoring - for this reason, it is common to restrict attention to pairs in which at least one individual was observed to fail and the ordering can be determined based on the information from the other member. Empirically, one can compute the proportion of pairs for which an ordering can be determined for which the pairs exhibit concordance in their observed and predicted ordering. To avoid bias that can arise from the sub-sample of pairs for which ordering is possible, estimators incorporating inverse probability of censoring weights have been developed.[59] More direct model-based estimates of concordance are also possible - see Gӧnen and Heller[60]. Blanche et al.[61] recently argued that when interest lies in predicting survival status at a time horizon t_o then computing the time-dependent area under the receiver operating characteristic curve is more suitable since it is aligned with the aim of predicting the dichotomous outcome (survived or not to t_o).

More extreme forms of censoring arise if it is necessary to examine individuals in order to determine whether they have experienced the event of interest or not. Clinical settings where this is the case include the evaluation of prediction models for asymptomatic vertebral compression in osteoporosis studies, prediction of the development of metastatic lesions in cancer patients with skeletal metastases, and prediction of progression of liver disease where the damage status is only measurable from biopsy. Interval-censoring arises when multiple assessments are made over time with the event time only known to have occurred between the last negative and the first assessment. Wu and Cook[62] discuss the development of predictive models in interval-censored data via penalization, with methods for assessing predictive accuracy addressing interval-censoring in the validation sample given in Wu and Cook[63]. Yang et al.[64] consider development of survival trees based on an extreme form of interval censoring yielding current status data; here, there is a single assessment time so observations are either left- or right-censored.

7. Concluding Remarks

There is an increasing awareness of the power of machine learning and artificial intelligence and it is natural that these are seeing application to problems of predictive inference. Ensemble methods can offer predictions that perform better than algorithms based on simple risk scores in many settings, and if there is no need to report easily calculable risk scores, ensemble and other “black-box" methods can often be adopted for best performance. Guidelines have been established for the steps appropriate for the development, evaluation and reporting of new predictive models.[65,66]

In many datasets used for the development of predictive model, there is relatively little attention paid to how individuals were recruited - recruitment mechanisms can be quite different for cohort studies, disease registries, and administrative datasets, and the details on how participants are selected can have a profound affect on the covariate and hence the risk score distributions. It is important to appreciate that the predictive performance of a predictive model will differ in populations with different covariate distributions and quantifying the extent to which performance deteriorates is part of the rationale for using external validation samples. If it is well-understood, and possible to model, how the training sample and target populations characteristics differ, calibration can be carried out. Otherwise, it is important to bear in mind that this may be a key factor in explaining poor transportability of predictive models to new settings.[67]

In the context of the linear model of Section 2, classical regression and causal inference were discussed as related topics, but ones that are distinct from predictive modeling. However, predictive modeling can play an important role in the development of propensity scores for causal analysis.[68,69] A newer connection between causal inference and predictive modeling arises in settings where interest lies in prediction of counterfactual outcomes to guide medical decision making.[70,71]

We have focused on continuous, binary and failure time responses, but there is great interest in the development of predictive models and methods for more general and complex life history processes in settings with incomplete data due to loss to follow-up. The frameworks we have discussed and the methods of measuring predictive accuracy can be adapted for such settings; we refer readers to Steyerberg[5] for a textbook treatment of this problem and Spitoni et al.[72] for remarks on prediction with multistate models; see also Section 8.2 of Cook and Lawless[73].

We close with some guiding principles to consider when analysing a research dataset, where the suitable approach is naturally shaped by the precise scientific aims.

1. If interest lies in assessing the relationship between a modest set of covariates and an outcome of interest, a traditional regression model can be fitted. Before interpreting the results of a fitted model, however, diagnostic checks should be carried out to assess model adequacy. Provided the underlying assumptions and model fit seem reasonable, covariate effects can be interpreted *in concert* - that is, the coefficient of a given variable reflects the association between that covariate and the response, while adjusting for the effects of the other covariates in the model.

2. If there is one covariate (such as an exposure or treatment) that is of primary interest, and the goal is to make a more formal assessment about the possible causal effect of that variable on the response, then careful thought is required about possible confounding variables to ensure they are adjusted for appropriately. The "potential outcomes" or counterfactual framework for causal reasoning has dominated the field of causal inference, which has seen significant high-impact advances over the past twenty to thirty years. Many of these methods exploit propensity scores derived from models of the exposure variable given the confounders. Propensity scores can be used as the basis for matching or stratification within a sample, or through regression adjustment; use of inverse propensity of exposure weighting is also particularly appealing, particularly when the covariate of interest is discrete.

3. Sometimes interest lies in identifying (i.e. selecting) which covariates in a large set of covariates appear to have an important relationship with the response. One approach for doing this is to carry out univariate tests of the association between each covariate and the response in turn, while controlling the risk of false positive findings through use of a multiple testing procedure. A second approach is to select the important covariates while considering them jointly - this is achieved through use of penalized regression models (e.g. the LASSO), which seeks to find the simplest set of covariates by penalizing an objective function for the accommodation of non-zero regression coefficients. The notion in this framework is that covariates exhibit a strong enough effect on the response that the improvement in fit outweighs the penalty for the inclusion of another covariate. The different penalty functions available for penalized regression operate in this way but differ in the nature and extent of the penalties for the use of more complex models.

4. If interest lies in predicting the response, the concerns regarding over-fitting and transportability arise. Specified (i.e. fixed) lists of covariates can be used to develop predictive models, but risk scores and predictive models will be optimized for the training dataset and some loss of performance typically expected in a validation sample. Often, some degree of model selection is carried out when building a predictive model (e.g. step-wise regression, penalized regression) since it may be desirable to have a small list of key variables when developing risk scores for prediction. When such procedures are employed, it is important to

note that use of standard estimators and their standard errors do not yield valid inference; the standard approach to constructing confidence intervals or reporting p-values are not valid in settings where a model was obtained following any variable selection technique. So-called ensemble methods of prediction recognize the instability of estimates from particular selected models, and the fact that many models may provided good fit to a given dataset. By synthesizing predictions across candidate models, predictive accuracy can be enhanced but again this is achieved at the price of not having interpretable point estimates of individual covariate effects.

With the increasing complexity of data in the information age, statistical methods are playing an increasingly important role in health and medical research. A clear specification of the primary scientific objectives is needed to identify the most suitable methods for analysis to address the questions of interest.

Funding

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849 and RGPIN 04207 for RJC) and the Canadian Institutes of Health Research (FRN 13887 for RJC).

References

1. Hernán M.A., Robins J.M. *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC 2020.

2. Greenland S., Pearl J., Robins J.M. Causal diagrams for epidemiologic research. *Epidemiology.* 1999; 10: 37–48.

3. Chan I.S., Ginsburg G.S. Personalized medicine: progress and promise. *Annual Review of Genomics and Human Genetics.* 2011; 12: 217–244.

4. Grant S.W., Collins G.S., Nashef S.A.M. Statistical primer: developing and validating a risk prediction model. *European Journal of Cardio-Thoracic Surgery.* 2018; 54: 203–208.

5. Steyerberg E.W. *Clinical Prediction Models: A Practical Approach to Development, Validation and Updating, Second Edition*. Cham, Switzerland: Springer Nature Switzerland AG 2019.

6. Byon H.-J., Lim C.-W., Lee J.-H., et al. Prediction of fluid responsiveness in mechanically ventilated children undergoing neurosurgery. *British Journal of Anaesthesia.* 2013; 110: 586–591.

7. Chen W., Fong J.W.H., Lind C.R.P., Knuckey N.W. P-POSSUM scoring system for mortality prediction in general neurosurgery. *Journal of Clinical Neuroscience.* 2010; 17: 567–570.

8. Copeland G.P., Jones D., Walters M. POSSUM: a scoring system for surgical audit. *British Journal of Surgery.* 1991; 78: 355–360.

9. Kent D.M., Paulus J.K., Sharp R.R., Hajizadeh N. When predictions are used to allocate scarce health care resources: three considerations for models in the era of Covid-19. *Diagnostic and Prognostic Research.* 2020; 4: 1–3.

10. Alba A.C., Agoritsas T., Walsh M., et al. Discrimination and calibration of clinical pre- diction models: users' guides to the medical literature. *Journal of the American Medical Association.* 2017; 318: 1377–1384.

11. Lo B.W.Y., Fukuda H., Angle M., et al. Clinical outcome prediction in aneurysmal sub-arachnoid hemorrhage: Alternations in brain-body interface. *Surgical Neurology International.* 2015; 7: S527–S537.

12. Owen S.V., Froman R.D. Uses and abuses of the analysis of covariance. *Research in Nursing & Health.* 1998; 21: 557–562.

13. Schemper M. Predictive accuracy and explained variation. *Statistics in Medicine.* 2003; 22: 2299–2308.

14. Harrell F.E., Califf R.M., Pryor D.B., Lee K.L., Rosati R.A. Evaluating the yield of medical tests. *Journal of the American Medical Association.* 1982; 247: 2543–2546.

15. Dobson A.J., Barnett A.G. *An Introduction to Generalized Linear Models*. Boca Raton, FL: CRC Press 2018.

16. Cook R.J. Generalized linear models in *Biostatistics and Genetic Epidemiology* (Elston R., Olson J., Palmer L., eds.), Chichester, West Sussex, UK: John Wiley & Sons Ltd 2002.

17. McFadden D. Conditional logit analysis of qualitative choice behavior in *Frontiers in Econometrics* (Zarembka P. , ed.) ch. 4, : 105–142, New York, NY: Academic Press 1974.

18. Nagelkerke N.J.D. A note on a general definition of the coefficient of determination. *Biometrika.* 1991; 78: 691–692.

19. Menard S. Coefficients of determination for multiple logistic regression analysis. *The American*

Statistician. 2000; 54: 17–24.

20. Rufibach K. Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology.* 2010; 63: 938–939.

21. Redelmeier D.A., Bloch D.A., Hickam D.H. Assessing predictive accuracy: how to compare Brier scores. *Journal of Clinical Epidemiology.* 1991; 44: 1141–1146.

22. Gneiting T., Balabdaoui F., Raftery A.E. Probability forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Methodological).* 2007; 69: 243–268.

23. Hosmer D.W., Hosmer T., Le Cessie S., Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine.* 1997; 16: 965–980.

24. Friedman J., Hastie T., Tibshirani R. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Science + Business Media, 2001.

25. Pepe M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press 2003.

26. Pencina M.J., D'Agostino R.B. Evaluating discrimination of risk prediction models: the C-statistic. *Journal of the American Medical Association.* 2015; 314: 1063–1064.

27. Van Calster B., Van Belle V., Vergouwe Y., Steyerberg E.W. Discrimination ability of prediction models for ordinal outcomes: relationships between existing measures and a new measure. *Biometrical Journal.* 2012; 54: 674–685.

28. Kraft P. Curses - winner's and otherwise - in genetic epidemiology. *Epidemiology.* 2008; 19: 649–651.

29. Zhong H., Prentice R.L. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics.* 2008; 9: 621–634.

30. Hoerl A.E., Kannard R.W., Baldwin K.F. Ridge regression: some simulations. *Communications in Statistics – Theory and Methods.* 1975; 4: 105–123.

31. Powers S., McGuire V., Bernstein L., Canchola A.J., Whittemore A.S. Evaluating disease prediction models using a cohort whose covariate distribution differs from that of the target population. *Statistical Methods in Medical Research.* 2019; 28: 309–320.

32. Breiman L. Heuristics of instability and stabilization in model selection. *Annals of Statistics.*

1996; 24: 2350–2383.

33. Hoerl A.E., Kannard R.W. Ridge regression: applications to nonorthogonal problems. *Technometrics.* 1970; 12: 55–67.

34. Tibshirani R.. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B.* 1996; 58: 267–288.

35. Fan J., Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association.* 2001; 96: 1348–1360.

36. Zou H., Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics.* 2008; 36: 1509–1533.

37. Zou H. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association.* 2006; 101: 1418–1429.

38. Zou H., Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B.* 2005; 67: 301–320.

39. Yuan M., Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B.* 2006; 68: 49–67.

40. Zhang C.H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics.* 2010; 38: 894–942.

41. Cox D.R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B.* 1972; 34: 187–202.

42. Mingers J. An empirical comparison of selection measures for decision-tree induction. *Machine Learning.* 1989; 3: 319–342.

43. Breiman L., Friedman J.H., R.A. Olshen, Stone C.J. *Classification and Regression Trees.* Boca Raton, FL: Chapman and Hall/CRC 1984.

44. Strobl C., Malley J., Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods.* 2009; 14: 323–348.

45. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria 2021.

46. Friedman J., Hastie T., Tibshirani R. Regularization paths for generalized linear models via

coordinate descent. *Journal of Statistical Software.* 2010; 33: 1–22.

47. Therneau T., Atkinson B. *rpart: Recursive Partitioning and Regression Trees* 2019. R package version 4.1-15.

48. Robin X., Turck N., Hainard A., et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011; 12: 77.

49. Liaw A., Wiener M. Classification and Regression by randomForest. *R News.* 2002; 2: 18– 22.

50. Andri et al. Signorell. *DescTools: Tools for Descriptive Statistics* 2021. R package version 0.99.42.

51. Tibshirani R. The LASSO method for variable selection in the Cox model. *Statistics in Medicine.* 1997; 16: 385–395.

52. LeBlanc M., Crowley J. Relative risk trees for censored survival data. *Biometrics.* 1992; 48: 411–425.

53. LeBlanc M., Crowley J. A review of tree-based prognostic models in *Recent Advances in Clinical Trial Design and Analysis. Cancer Treatment and Research, Vol 75* (Thall P.F., ed.): 113–124, Boston, MA: Springer 1995.

54. Molinaro A.M., Dudoit S., Van der Laan M.J. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis.* 2004; 90: 154–177.

55. Steingrimsson J.A., Diao L., Strawderman R.L. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association.* 2019; 114: 370–383.

56. Zhu M., Fan G. Variable selection by ensembles for the Cox model. *Journal of Statistical Computation and Simulation.* 2011; 81: 1983–1992.

57. Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S. Random survival forests. *The Annals of Applied Statistics.* 2008; 2: 841–860.

58. Harrell Jr F.E., Lee K.L., Mark D.B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine.* 1996; 15: 361–387.

59. Uno H., Cai T., Pencina M.J., D’Agostino R.B., Wei L.J. On the C-statistics for evaluating overall

adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine.* 2011; 30: 1105–1117.

60. G ̈onen M., Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika.* 2005; 92: 965–970.

61. Blanche P., Kattan M.W., Gerds T.A. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics.* 2019; 20: 347–357.

62. Wu Y., Cook R.J. Penalized regression for interval-censored times of disease progression: Selection of HLA markers in psoriatic arthritis. *Biometrics.* 2015; 71: 782–791.

63. Wu Y., Cook R.J. Assessing the accuracy of predictive models with interval-censored data. *Biostatistics.* 2020: kxaa011.

64. Yang C., Diao L., Cook R.J. Survival trees for current status data in *Survival Prediction-Algorithms, Challenges and Applications*. 2021: 83–94.

65. Collins G.S., Reitsma J.B., Altman D.G., Moons K.G.M. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *British Journal of Surgery.* 2015; 102: 148–158.

66. Moons K.G., Altman D.G., Reitsma J.B., et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine.* 2015; 162: W1–W73.

67. Justice A.C., Covinsky K.E., Berlin J.A. Assessing the generalizability of prognostic information. *Annals of Internal Medicine.* 1999; 130: 515–524.

68. Lee B.K., Lessler J., Stuart E.A. Improving propensity score weighting using machine learning. *Statistics in Medicine.* 2010; 29: 337–346.

69. Zhu Y., Schonbach M., Coffman D.L., Williams J.S. Variable selection for propensity score estimation via balancing covariates. *Epidemiology.* 2015; 26: e14–e15.

70. van Geloven N., Swanson S.A., Ramspek C.L., et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology.* 2020; 35: 619–630.

71. Prosperi M., Guo Y., Sperrin M., et al. Causal inference and counterfactual prediction in

machine learning for actionable healthcare. *Nature Machine Intelligence.* 2020; 2: 369–375.

72. Spitoni C., Lammens V., Putter H. Prediction errors for state occupation and transition probabilities in multi-state models. *Biometrical Journal.* 2018; 60: 34–48.

73. Cook R.J., Lawless J.F. *Multistate Models for the Analysis of Life History Data*. Boca Raton, FL: CRC Press 2018.