Simplifying a prognostic model: a simulation study based on clinical data

Gareth Ambler¹,*,†, Anthony R. Brady² and Patrick Royston³

¹Department of Statistical Science, University College, 1-19 Torrington Place, London WC1E 7HB, U.K.
²Intensive Care National Audit and Research Centre, Tavistock House, Tavistock Square, London WC1H 9HR, U.K.
³MRC Clinical Trials Unit, 222 Euston Road, London NW1 2DA, U.K.

SUMMARY

Prognostic models are designed to predict a clinical outcome in individuals or groups of individuals with a particular disease or condition. To avoid bias many researchers advocate the use of full models developed by prespecifying predictors. Variable selection is not employed and the resulting models may be large and complicated. In practice more parsimonious models that retain most of the prognostic information may be preferred. We investigate the effect on various performance measures, including mean square error and prognostic classification, of three methods for estimating full models (including penalized estimation and Tibshirani’s lasso) and consider two methods (backwards elimination and a new proposal called stepdown) for simplifying full models. Simulation studies based on two medical data sets suggest that simplified models can be found that perform nearly as well as, or sometimes even better than, full models. Optimizing the Akaike information criterion appears to be appropriate for choosing the degree of simplification. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: prognostic models; variable selection; penalisation; lasso; ROC; AIC

1. INTRODUCTION

Prognostic models are designed to predict a clinical outcome in individuals or group of individuals with a particular disease or condition. For example, in oncology the aim may be to estimate the probability of a patient surviving a given time following a cancer diagnosis.

* Correspondence to: Gareth Ambler, Department of Statistical Science, University College, 1-19 Torrington Place, London WC1E 7HB, U.K.
† E-mail: g.ambler@ucl.ac.uk

Contract/grant sponsor: The Wellcome Trust; contract/grant number: 045512/Z/95/Z.
Contract/grant sponsor: BUPA Foundation.

Received January 2002
Accepted October 2002
At diagnosis information is usually collected for each individual; in the case of breast cancer this may include menopausal status, tumour grade and the number of positive lymph nodes as well as standard demographic variables such as age. The statistical problem is to produce a model based on this information that produces accurate predictions for new patients or patient groups with the same disease or condition.

A large number of prognostic models appear in the medical and statistical literature. However, many of these models either have not been tested on or fail to work satisfactorily for new patients. These models have not been validated [1]. There are a number of reasons why a model may fail to work in new data. There may have been insufficient data in the original sample to estimate the parameters reliably or a lack of intrinsic prognostic information in the candidate predictors [1]. Simplification procedures may have introduced both selection and omission bias by selecting only predictors with large observed effects [2, 3]. Model misspecification may have resulted in a loss of information and the introduction of bias [4]; examples of misspecification include modelling curved relationships as linear and dichotomizing continuous variables. Also the model may fail to work because of geographical and temporal differences in clinical practice; a model developed at one location may not work at another location or even at the same location a few years later.

Since search algorithms, such as stepwise selection of variables, introduce bias, several authors have suggested that they should not be employed and models should be prespecified. Harrell claims that this approach provides more reliable prognostic models and has suggested guidelines for developing such models [2]. As well as model prespecification the guidelines suggest that the complexity of the model should be limited only by the number of events in the data. Such models are often described as full models since they may be viewed as the largest models permitted by the data.

A full model may be large and complicated, particularly if it is developed using a data set with many observations and potential predictors. In practice a more parsimonious model is preferable since predictors are often difficult or expensive to measure reliably. Clinicians and research workers may also be discouraged by a large prognostic model since its practical application may be time-consuming and complicated [5].

The aim of this paper is to investigate techniques for estimating and simplifying full models without sacrificing much prognostic ability. We aim to produce reduced models that contain fewer parameters and predictors but contain the important relationships. This strategy may be used to screen all clinically relevant variables by including them in the full model, one of the prerequisites for producing models that are clinically credible [5].

We consider estimating full models using three different techniques: maximum likelihood (ML); penalized ML [6], and the lasso [7]. The first method will produce a better fit to data but penalized estimation and the lasso (itself a form of penalized estimation) introduce shrinkage that biases the regression coefficients towards zero but makes predictions that are claimed to have reduced mean squared error.

Two types of simplification strategy are considered. The first uses the stepdown approach suggested by Harrell [2]. A full model is fitted and the prognostic index (linear predictor) obtained. Predictors which have the weakest relationship with this index are dropped and the remaining predictors are used to approximate the index. The method is similar to backwards elimination, our second strategy, but the original outcomes are only used to fit the full model: Harrell and colleagues claim that this strategy avoids overfitting since any penalization remains in the linear predictor [2].
We investigate the estimation and simplification strategies using simulated data derived from two medical data sets. The first is from a study of the treatment of breast cancer. We consider the binary outcome of recurrence-free survival for 3 years. The second is from a study of patients with abdominal aortic aneurysms, with a binary outcome of 3 year survival.

The structure of the paper is as follows. Section 2 describes the two medical data sets. Section 3 describes the design of the study and the methods used for estimation, simplification and assessment of prognostic performance. Sections 4 and 5 describe, respectively, simulation details and results. Section 6 is a discussion.

2. DATA SETS

2.1. Breast cancer data

The data set consists of a ‘comprehensive cohort’ of 720 patients with primary node positive breast cancer who were recruited by the German Breast Cancer Study Group (GBSG) between July 1984 and December 1989 [8]. Sauerbrei and Royston [4] used the data to demonstrate the fractional polynomial approach [9] to building prognostic models for a survival time outcome with several predictors, some continuous and some binary. The majority of the data (686 patients) are complete for recurrence-free survival time and for the continuous predictors age, tumour size, number of positive lymph nodes, progesterone receptor status and oestrogen receptor status and the categorical predictors menopausal status, tumour grade and hormone treatment. Menopausal status and treatment are binary. Tumour grade has three ordered levels.

For the purposes of this study we chose to define a binary outcome, 3 year recurrence-free survival. Data for individuals who were censored before 3 years were imputed by fitting a log-normal survival model to the original data and randomly generating new survival times for the censored observations from the estimated distribution [10]. After imputation we had 249 (36 per cent) events. In practice we would use survival analysis to analyse these data since the 3 year cut-off is arbitrary and the use of logistic regression is statistically inefficient.

We fitted a multivariable logistic model to the data using ML to investigate the importance of each predictor: Table I describes the effect that the omission of each predictor had on the model. Sauerbrei and Royston [4] detected curved relationships with several of the continuous predictors so we used regression splines with 4 d.f. (number of positive lymph nodes only has three due to its highly skewed distribution) to model the continuous predictors (see Section 3). Tumour grade was modelled using two binary variables. Number of nodes was the most important predictor in the model, followed by age and number of progesterone receptors. Menopausal status was highly correlated with age (Spearman’s \( \rho = 0.80 \)) and had a weak relationship with the response after adjusting for all other predictors. The next highest correlation was between progesterone receptor status and oestrogen receptor status (\( \rho = 0.60 \)).

2.2. Surgical data

A total of 2305 patients with abdominal aortic aneurysm (AAA) were recruited into the UK Small Aneurysm Trial and Study between 1991 and 1995 [11]. All patients were flagged at the Office of National Statistics to enable automatic notification of emigration or death. At the end of the study period (July 1999) follow-up for all-cause mortality was complete up to 3.6 years. We chose to classify patients according to 3 year survival status, with 514 (22 per cent) events
Table I. Statistical significance of the predictors in a multivariable model fitted (using ML) to the breast cancer data set. Continuous predictors are modelled using regression splines. $\chi^2$ denotes the decrease in the likelihood ratio $\chi^2$ statistic for the model when the predictor is omitted and the model re-fitted.

The likelihood ratio $\chi^2$ statistic for the model with all predictors is 153.3.

<table>
<thead>
<tr>
<th>Variable</th>
<th>D.F.</th>
<th>$\chi^2$</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of positive lymph nodes</td>
<td>3</td>
<td>60.8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>4</td>
<td>15.7</td>
<td>0.008</td>
</tr>
<tr>
<td>Progesterone receptor status</td>
<td>4</td>
<td>14.4</td>
<td>0.008</td>
</tr>
<tr>
<td>Hormone therapy</td>
<td>1</td>
<td>9.6</td>
<td>0.002</td>
</tr>
<tr>
<td>Oestrogen receptor status</td>
<td>4</td>
<td>7.6</td>
<td>0.15</td>
</tr>
<tr>
<td>Tumour size</td>
<td>4</td>
<td>2.4</td>
<td>0.67</td>
</tr>
<tr>
<td>Tumour grade</td>
<td>2</td>
<td>2.3</td>
<td>0.33</td>
</tr>
<tr>
<td>Menopausal status</td>
<td>1</td>
<td>1.0</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table II. Statistical significance of the predictors in a multivariable model fitted (using ML) to the aortic aneurysm data set. $\chi^2$ denotes the decrease in the likelihood ratio $\chi^2$ statistic for the model when the predictor is omitted and the model re-fitted. The likelihood ratio $\chi^2$ statistic for the model with all predictors is 222.8. All the predictors (excluding the constant) have been standardized to have zero mean and unit variance.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\beta}$</th>
<th>$\chi^2$</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(AAA diameter)</td>
<td>0.35</td>
<td>42.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ABPI</td>
<td>-0.24</td>
<td>18.4</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IHD</td>
<td>0.20</td>
<td>14.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.20</td>
<td>13.2</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>BMI</td>
<td>-0.20</td>
<td>12.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FEV1</td>
<td>-0.21</td>
<td>12.1</td>
<td>0.001</td>
</tr>
<tr>
<td>HT drug</td>
<td>0.20</td>
<td>12.1</td>
<td>0.001</td>
</tr>
<tr>
<td>Log(creatine)</td>
<td>0.17</td>
<td>9.2</td>
<td>0.003</td>
</tr>
<tr>
<td>Age</td>
<td>0.15</td>
<td>7.0</td>
<td>0.008</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.07</td>
<td>2.3</td>
<td>0.13</td>
</tr>
<tr>
<td>Log(WCC)</td>
<td>0.07</td>
<td>1.7</td>
<td>0.20</td>
</tr>
<tr>
<td>SBP</td>
<td>0.07</td>
<td>1.1</td>
<td>0.31</td>
</tr>
<tr>
<td>Aspirin</td>
<td>0.02</td>
<td>0.1</td>
<td>0.74</td>
</tr>
<tr>
<td>DBP</td>
<td>0.01</td>
<td>0.1</td>
<td>0.83</td>
</tr>
<tr>
<td>Male</td>
<td>-0.01</td>
<td>0.0</td>
<td>0.91</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.41</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

for this outcome. A set of 15 candidate predictors of survival were specified before analysis of the data based on published studies and clinical knowledge of the investigators. These were age, sex, ankle-brachial pressure index (ABPI, a measure of atherosclerosis), forced expiratory volume in 1s (FEV1), body mass index (BMI), evidence of ischaemic heart disease (IHD) on electrocardiogram, diabetes, systolic and diastolic blood pressure (SBP and DBP), use of aspirin, use of anti-hypertensive drugs (HT drugs), current smoking, log transformed AAA diameter, log transformed blood creatinine and log transformed white cell count (WCC).

A logistic model was fitted to the data using ML (Table II). We used linear terms for all predictors as suggested by a previous unpublished analysis of the data. It is apparent
that, in the multivariable model, there was one very strong predictor, eight moderate predictors and six weak predictors. Generally there were only modest correlations in the data set (Spearman $\rho < 0.24$), the exceptions being between SBP and DBP ($\rho = 0.62$), and FEV$_1$ and male sex ($\rho = 0.37$).

3. METHODS

3.1. Design of study

We now give a brief overview of the study design. For each of the two data sets described in Section 2 we derived a true model and specified a full model. From the true model we obtained a set of true probabilities which were used to generated new responses and hence new (simulated) data sets. These were used to evaluate the following factors:

(i) method of full model estimation – ML, penalized ML or lasso;
(ii) method of full model simplification – stepdown or backwards elimination (with ML fit only) at various levels;
(iii) size of data set – two sizes were considered for both data sets;
(iv) presence or otherwise of uncorrelated, uninfluential variables – with aortic aneurysm data only.

For each combination of factors we performed 1000 simulations. The performance of each estimation and simplification strategy was assessed by comparing the fitted models to both the simulated data and the true model. The following criteria were used depending on the comparison:

(i) true model – prognostic classification and mean squared error;
(ii) data – Brier score and concordance.

In the rest of this section we describe the methods used to specify, estimate and simplify the full models, and the performance criteria we used. In Section 4 we discuss data-related issues including derivation of the true models.

3.2. Specification

The full model approach requires us to prespecify the predictors that will appear in the model and the overall functional form of the model [2, 12]. The choice of predictors often depends on existing knowledge and biological plausibility. The form of the model concerns the complexity of any predictor transformations, the inclusion of interaction terms and the metric for predictor effects; in this paper we use logistic additive models without interactions. A consequence of prespecifying the model is that confidence intervals and hypothesis tests are statistically valid, at least asymptotically.

There may be important curved relationships in the data that should be modelled accurately to reduce bias and hence improve the accuracy of the model’s predictions. For example, other investigators have shown that the breast cancer data, in the context of the Cox model, contain several predictors that have strong curved relationships with the response [4]. Consequently
we prespecify cubic regression splines to model the relationship between each continuous predictor and the response [13, 14]. Briefly, a cubic regression spline is a set of piecewise cubic polynomials joined together at points known as knots. The spline is often constrained to be continuous at the knots, have continuous first and second derivatives and be constrained to be linear beyond two boundary knots placed at the extremes of the predictor values. A spline based on \( K \) knots (excluding boundary knots) has \( K + 1 \) terms and hence \( K + 1 \) adjustable parameters (d.f.). A common choice for the placement of knots is at the quantiles of the predictor values, for example a spline with four parameters would have three knots placed at the 25 per cent, 50 per cent and 75 per cent quantiles as well as boundary knots at the 0 per cent and 100 per cent quantiles.

Several authors make recommendations concerning the maximum number of predictors (parameters) that should be estimated in a prognostic model. A common opinion is that ratio of events to predictors should not be less than 10:1 where, in the binary data setting, the number of events is defined as either the number of failures (1s) or the number of successes (0s), whichever is the lower [2, 15, 16]. This ratio is called the events per variable (EPV). Steyerberg et al. suggest that shrinkage should be used if the EPV is less than 10 and should be considered if the EPV is less than 20 in prespecified models [16].

### 3.3. Estimation

As well as standard ML estimation, we consider two types of penalized estimation. The first is the penalized approach described by Le Cessie and van Houwelingen [6]. A penalty is introduced into the likelihood formula so that the objective function becomes

\[
\log L - \frac{1}{2} \lambda \beta^T P \beta
\]

where \( L \) is the usual likelihood function, \( \beta \) the vector of parameters (excluding the intercept) and \( \lambda \) a scalar fixed by the investigator. The diagonal matrix \( P \), whose elements are the variances of the predictors in the model, ensures that the penalty is invariant to the scaling of the predictors.

The main issue is the choice of \( \lambda \). For any \( \lambda > 0 \), shrinkage is introduced and the effective degrees of freedom (e.d.f.) of the model is reduced. We may estimate e.d.f. (see Gray [17]) and calculate a modified AIC statistic [18]

\[
AIC_{\text{mod}} = 2 \log L - 2 \times \text{e.d.f.}
\]

Using a grid search we choose \( \lambda \) to maximize \( AIC_{\text{mod}} \) since this value optimizes a cross-validation statistic, a measure of predictive accuracy [2].

The second type of estimation we consider is the lasso (least absolute shrinkage and selection operator) [7]. This maximizes \( \log L \) subject to the constraint

\[
\sum_j |\beta_j| \leq t
\]

where \( \beta_j \) is the \( j \)th parameter estimate and \( t \) is a shrinkage parameter chosen by the investigator. Tibshirani [7] derives an expression for e.d.f. and suggests choosing \( t \) to minimize the generalized cross-validation style statistic

\[
\text{GCV} = \frac{-\log L}{n(1 - \text{e.d.f.}/n)^2}
\]
another measure of predictive accuracy. For each value of $t$ we may calculate

$$ s = \frac{\sum_j |\hat{\beta}_j^{lasso(t)}|}{\sum_j |\hat{\beta}_j^{ML}|} $$

which quantifies the amount of shrinkage introduced, where $\hat{\beta}_j^{lasso(t)}$ and $\hat{\beta}_j^{ML}$ are the lasso and ML estimates, respectively. In practice a grid search is performed over $s$ to minimize GCV [7]. A potential advantage of lasso is that variable selection is an integral part of the estimation process; unlike the penalized approach, lasso tends to produce some parameter estimates that are exactly 0, especially for lower values of $s$ [7]. However, even when some parameter estimates are exactly 0, we still think of the lasso model as a full model since no terms are excluded from the linear predictor $a$ priori.

For convenience we refer to a full model fit using ML as FM. Similarly, full model fits using penalized ML and lasso are referred to as FPM and FLM, respectively.

### 3.4. Simplification

#### 3.4.1. Stepdown.

Stepdown may be used to approximate a prognostic index (linear predictor), where the prognostic index is simply a linear combination of the predictors in the model [2]. An (OLS) regression of the prognostic index on these predictors produces a perfect fit with an $R^2$ of 1. The index may be simplified by dropping the predictor whose omission causes the smallest decrease in $R^2$. This procedure continues until the omission of any more predictors would result in an $R^2$ lower than a prespecified level, say 0.95. This algorithm results in the omission of predictors that have a small effect on the prognostic index. If all the predictors in the model are uncorrelated and standardized, it is straightforward to show (see Appendix) that the omission of the $k$th predictor $X_k$ results in a decrease in $R^2$ given by

$$ \Delta R^2_k = \frac{\hat{\beta}_k^2}{\sum_j \hat{\beta}_j^2} $$

where $\hat{\beta}_j$ is the $j$th parameter estimate from the full model. Hence predictors with small coefficients, after allowing for scaling, will have little effect on $R^2$. With correlated predictors $\Delta R^2_k$ will depend on correlations between predictors in the model and predictors that have already been omitted.

We use stepdown to approximate FM, FPM and FLM and consider the $R^2$ values $\{0.99, 0.95, 0.90, 0.80, 0.70, 0.50\}$. In the case of lasso, some predictors may already have been dropped, but we still consider stepdown as a way of further simplifying the model. We use the notation FM($R^2$), FPM($R^2$) and FLM($R^2$) to refer to full model fits approximated using stepdown at the $R^2$ level.

#### 3.4.2. Backwards elimination.

Backwards elimination (BE) typically starts from a model containing several terms then drops those that are not statistically significant at a nominal level $\alpha$, often 0.05. At each stage the term that is least significant is dropped; the process continues until there are no non-significant terms to drop. This technique differs from stepdown in that a new model is fitted to the original data at each stage; with stepdown the original response
is only used to fit the full model. Also BE is generally only used in conjunction with ML estimation whereas stepdown may be applied to any prognostic index regardless of how it was obtained.

We consider the significance levels \{0.50, 0.30, 0.157, 0.1, 0.05, 0.01, 0.001\} and use the notation \(\text{BE}(\alpha)\) to refer to the model fitted using BE at the \(\alpha\) level. We note that the level 0.157 selects the predictors that are usually chosen by maximizing the AIC in all-subsets procedure (if all predictors have 1 d.f.) [19].

In total we consider three full models, FM, FPM and FLM, and four types of simplified models, \(\text{FM}(R^2)\), \(\text{FPM}(R^2)\), \(\text{FLM}(R^2)\) and \(\text{BE}(\alpha)\). We note that FM is equivalent to both \(\text{FM}(1)\) and \(\text{BE}(1)\). Likewise FPM and FLM are equivalent to \(\text{FPM}(1)\) and \(\text{FLM}(1)\), respectively.

3.4.3. Multiple terms. Often the full model will contain multiple terms linked to a single predictor. This may occur if we have categorical variables with more than two levels, or we use techniques such as regression splines and fractional polynomials. Two strategies are to consider either all the linked terms together or every term individually. We investigate the first strategy since our primary interest is the inclusion or omission of variables rather than simplification of their functional forms; the inclusion of a variable in a model determines that future users of the model have to measure the variable. With stepdown the decrease in \(R^2\) upon dropping all linked terms is computed. With BE the significance of these multiple terms using a likelihood ratio test is considered.

3.4.4. AIC strategy. We investigate several values of \(R^2\) for stepdown and several significance levels for BE. We are able to assess their performance since we know the true model. However, in practice the true model is unknown and it may not be clear which simplified model offers an acceptable compromise between accuracy and parsimony. To automate the procedure, we investigate an approach in which the AIC statistic

\[
\text{AIC} = 2 \log L - 2p
\]

is maximized, where \(p\) is the number of parameters in the model. This statistic differs from \(\text{AIC}_{\text{mod}}\) in (1) where e.d.f. is substituted for \(p\).

We use this approach to choose separate \(R^2\) levels for FM, FPM and FLM and a significance level for the BE procedure. For stepdown, candidate models are the full model itself and those selected at the various \(R^2\) values. Each candidate model yields fitted probabilities which are used to calculate the binomial likelihood \(L\) and hence AIC. The chosen model is that with the highest AIC. For BE, the candidate models are the full model and those selected at the various significance levels. We anticipate that \(\text{BE}(0.157)\) will usually be chosen.

We use the notation \(\text{FM}(\text{AIC})\), \(\text{FPM}(\text{AIC})\), \(\text{FLM}(\text{AIC})\) and \(\text{BE}(\text{AIC})\) to denote models chosen using this strategy.

3.5. Performance criteria

We describe these criteria in the context of a single simulated data set. We denote the true probabilities, simulated responses and fitted probabilities by \(p_i\), \(y_{i}^{(\text{sim})}\) and \(p_{i}^{(\text{sim})}\), respectively, where \(i = 1, \ldots, n\).
3.5.1. True model. We use two measures to compare a fitted model with the underlying true model. The first, *agree*, is a measure of prognostic classification and is applied when we have categorized patients into risk groups based on their fitted probabilities. For example, we categorized patients in the breast cancer data set into five groups based on the cutpoints 0.2, 0.4, 0.6 and 0.8. *Agree* is simply the proportion of patients whose estimated and true risk groups match.

Our second performance measure is the mean square error (MSE) between the estimated and true probabilities. That is

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (p_i - \hat{p}_i^{(sim)})^2$$

We expect MSE to be lower for those models estimated using penalization.

3.5.2. Data. We also use two measures to compare a fitted model with (simulated) data. The first, concordance (*c*), quantifies the strength of the rank correlation between responses and fitted probabilities [20] and is commonly used as a measure of discrimination [2]. To calculate *c* we consider all possible pairs of responses where one subject failed (1) and the other did not (0). Concordance is then defined as the proportion of such pairs where the failure had the higher fitted probability. A value of 1 indicates perfect separation between those who failed and those who did not, whereas a value of 0.5 indicates random predictions [2].

The second performance measure, the Brier score, is the mean square difference between the responses and the estimated probabilities [2]. That is

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^{n} (y_i^{(sim)} - \hat{p}_i^{(sim)})^2$$

4. SIMULATIONS

4.1. Breast cancer

We based the true model on model III of Sauerbrei and Royston [4]. This (Cox) model was derived by specifying a negative exponential term for the number of positive lymph nodes and fractional polynomials (FPs) for the other continuous predictors. A multivariable FP algorithm was used to fit the model [22]. We used this approach and obtained our true model with four predictors (Figure 1).

We used the same prespecified full model as that described in Section 2. Regression splines with 4 d.f. were used to model the continuous predictors, and binary variables were used for the categorical predictors. Since all eight predictors were included, the full model contained four predictors not present in the true model. The model had 23 unknown parameters, excluding the intercept.

As mentioned in Section 3.1, we also investigated the effect of reducing the data set size to see how EPV affected the performance of the estimation and simplification strategies. Subsamples were taken from each full size data set with the EPV kept constant by sampling separately from failures (1) and successes (0). Two scenarios were investigated (see Table III), one involving full size data sets (scenario 1), the other involving subsets (scenario 2). The
Figure 1. Functional relationships of the predictors used to simulate the breast cancer data. The $y$-scale is log-odds. The $y$-axis for $f(\text{age})$ is truncated at 2 to enhance readability; the lowest value of age in the data is 21.

Table III. Simulation scenarios describing the sample size ($n$), number of events ($e$), number of true predictors (true), number of candidate predictors (candidate) and EPV.

<table>
<thead>
<tr>
<th>Data</th>
<th>Scenario</th>
<th>$n$</th>
<th>$e$</th>
<th>True</th>
<th>Candidate</th>
<th>EPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>1</td>
<td>686</td>
<td>249</td>
<td>4</td>
<td>4+4 (23 df)</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>317</td>
<td>115</td>
<td>4</td>
<td>4+4 (23 df)</td>
<td>5.0</td>
</tr>
<tr>
<td>Aortic aneurysm</td>
<td>1</td>
<td>2305</td>
<td>514</td>
<td>15</td>
<td>15</td>
<td>34.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2305</td>
<td>514</td>
<td>15</td>
<td>15+10</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>672</td>
<td>150</td>
<td>15</td>
<td>15</td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>672</td>
<td>150</td>
<td>15</td>
<td>15+10</td>
<td>6.0</td>
</tr>
</tbody>
</table>

fullsize data sets had an expected 249 events so the EPV was approximately 11. The subsets had 115 events ($n=317$) and an EPV of 5.

To establish a benchmark for the performance measures we fitted a correctly specified model (CSM) to each data set using ML, where the CSM had the same predictors as the true model with the correct functional forms. Finally, the cutpoints used to calculate agree were 0.2, 0.4, 0.6 and 0.8.
4.2. Aortic aneurysm

The true model for these data was derived by fitting a logistic model with all 15 predictors to the original data using ML (see Table II). In contrast to the small true model based on the breast cancer data, the aortic aneurysm derived true model has many predictors and incorporates a wide range of effect sizes.

With these data we also investigated the effect of including ten uninfluential, Normally distributed variables (noise variables) as candidate predictors. The addition of noise variables decreased the EPV and increased the number of variables considered for modelling. We considered four scenarios (Table III). Scenarios 1 and 2 used full size data sets ($n = 2305$) with an expected 514 events, whereas scenarios 3 and 4 used subsets ($n = 672$) with 150 events. Scenarios 2 and 4 included the noise variables.

The prespecified full model for scenarios 1 and 3 included all 15 predictors, whereas the full model for scenarios 2 and 4 included the 15 predictors plus the 10 noise variables. The candidate predictors were all modelled using linear terms so the full model had either 15 or 25 unknown parameters and was correctly specified in the former case. Therefore the EPVs for scenarios 1 and 2 were 34 and 21, respectively, and the EPVs for scenarios 3 and 4 were 10 and 6, respectively. Finally, the cutpoints used to calculate agree were 0.1, 0.2, 0.3 and 0.4.

5. RESULTS

5.1. Breast cancer

The results for scenario 1 are illustrated in Figure 2. Consider the solid lines representing $FM(R^2)$. The circles on these lines, from left to right, represent the simulations performed at $R^2$ values 0.70, 0.80, 0.90, 0.95, 0.99, respectively. Each circle shows the value of the performance measure (either agree, MSE, c or Brier) and the number of predictors, averaged over 1000 simulations, at that level of $R^2$. The average performances of $FM$ and $FM(AIC)$ are indicated on the lines by □ and △, respectively. The other lines follow the same pattern although the circles on the $BE(z)$ lines represent the $z$ values 0.001, 0.05, 0.157, 0.30, 0.50.

The lower MSE of both $FPM$ and $FLM$ is evident, although minor in this example. Of the full models $FLM$ provided the best agreement with the truth and had the lowest MSE, although $CSM$ achieved far better agree(0.83) and MSE(1.7) values (Table IV). This is partly due to bias, regression spline terms were used to model $FP$ and exponential terms, and partly due to the inclusion of uninfluential variables. When $FLM$ was simplified, its performance deteriorated. In contrast the opposite occurred for $FM$ and $FPM$ when assessing predictions from the simplified models with the true probabilities. Both agree and MSE achieved their best values after four predictors, on average, had been dropped. The automatic simplification (AIC) strategy tended to choose models with about four predictors that performed better than the corresponding full models.

Concordance was highest for $FM$, although the value for $FPM$ was similar. For all models $c$ decreased with increasing simplification, although only by about 0.01 after dropping four predictors. Brier was lowest for $FM$. The score increased with increasing simplification although little change can be seen until more than three predictors have been dropped. The $c$ and Brier values for $CSM$ (0.752 and 0.187, respectively, Table IV) were not as good as
Figure 2. Mean (a) agreement (b) MSE (c) concordance and (d) Brier score for FM ($R^2$), FPM ($R^2$), FLM ($R^2$) and BE ($\alpha$) at $R^2$ levels \{0.99, 0.95, 0.90, 0.80, 0.70\} and $\alpha$ levels \{0.50, 0.30, 0.157, 0.05, 0.001\} (the values are indicated from right to left on each line by $\bigcirc$). The full models are indicated by $\square$ on the relevant lines. Models selected by the automatic (AIC) strategy are indicated by $\triangle$. Simulated breast cancer data scenario 1 ($n = 686$, EPV = 11).

Table IV. Mean agree, MSE, concordance and Brier score of correctly specified models.

<table>
<thead>
<tr>
<th>Data</th>
<th>Scenario(s)</th>
<th>agree</th>
<th>MSE*</th>
<th>c</th>
<th>Brier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>1</td>
<td>0.83</td>
<td>1.7</td>
<td>0.752</td>
<td>0.187</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.78</td>
<td>3.2</td>
<td>0.756</td>
<td>0.185</td>
</tr>
<tr>
<td>Aortic aneurysm</td>
<td>1 and 2</td>
<td>0.77</td>
<td>1.3</td>
<td>0.711</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>3 and 4</td>
<td>0.62</td>
<td>4.2</td>
<td>0.726</td>
<td>0.152</td>
</tr>
</tbody>
</table>

* ×1000.

those achieved by all three full models. This reflects the optimism inherent in such data-based assessment measures when uninfluential predictors are fitted to the data and the measure is calculated on the same data used to estimate the parameters.

FLM obtained the best agree and MSE values in scenario 2 (Figure 3). Both FPM and FLM demonstrated substantially lower MSE and better agreement compared to FM in this low EPV scenario. The four models produced using the AIC strategy had similar values of agree, although FPM(AIC) and FLM(AIC) had lower values of MSE. The agree and MSE values for CSM were 0.78 and 3.2, respectively (Table IV). Of the models produced without penalization, BE($\alpha$) had lower MSE and higher agree values than FM($R^2$).
Figure 3. Mean (a) agreement (b) MSE (c) concordance and (d) Brier score for FM ($R^2$), FPM ($R^2$), FLM ($R^2$) and BE ($\alpha$) at $R^2$ levels {0.99, 0.95, 0.90, 0.80, 0.70} and $\alpha$ levels {0.50, 0.30, 0.157, 0.05, 0.001} (the values are indicated from right to left on each line by □). The full models are indicated by △ on the relevant lines. Models selected by the automatic (AIC) strategy are indicated by □. Simulated breast cancer data scenario 2 ($n = 317$, EPV = 5).

The plots for $c$ and Brier resemble those from scenario 1. Again CSM values ($c = 0.756$ and Brier = 0.185, Table IV) were inferior to those of the full models.

5.2. Aortic aneurysm

We first discuss the results for scenario 1 (Figure 4). All the strategies performed in a similar manner and the full models achieved a good agreement between predicted and true risk categories (0.77) as we might expect with a correctly specified model and a large data set. Simplification to less than about 11 predictors rapidly led to higher values of MSE and Brier and lower values of $c$ and $agree$. The AIC strategy resulted in models with similar performance values to the full models but with approximately four fewer variables on average.

When noise variables were added to the data sets (scenario 2) predictions from FLM were marginally closer to the true probabilities than FM as reflected by MSE and $agree$. However, FLM used an average of 5.8 fewer predictors than FM (Figure 5). Use of the AIC strategy resulted in models with performance equivalent to the full models, but with about 13 fewer variables on average. There was negligible difference between the four estimation and simplification strategies.

The results from scenarios 3 and 4 demonstrate the effects of reducing sample size, and hence EPV (Figures 6 and 7). Agreement between predicted and true risk categories was
reduced from 0.77 to 0.62 (FM with no noise variables) or 0.71 to 0.54 (FM with noise variables). These were reductions in agree of 19 per cent and 24 per cent respectively. Similarly MSE increased more than three-fold in both cases. However, concordance improved with lower EPV from 0.711 to 0.726 (FM with no noise variables) or 0.716 to 0.740 (FM with noise variables). The Brier score displayed similar improvement with reduced sample size. There was little difference between the full models in scenario 3 but when the noise variables were added FPM and FLM achieved better agree and MSE values than FM, although their c and Brier values were slightly inferior. In both scenarios the full models could be simplified to about 11 predictors with little loss in performance. In these low EPV scenarios, use of the AIC strategy resulted in a small loss of predictive accuracy.

6. DISCUSSION

The simulations allowed us to investigate the performance of full models in a variety of scenarios. When EPV was large (>20) there appeared to be little practical difference in predictive accuracy between the estimation methods even in the presence of noise variables. For lower values of EPV, performance seemed to depend on whether all the predictors were
influential or not. There was little difference between the estimation methods for the aortic aneurysm data where all 15 predictors were influential, but there were differences for the breast cancer data where only half the predictors were influential. FLM performed better than FPM and FM in the latter scenario. For the lowest values of EPV FLM performed the best, suggesting that this technique is useful when we have either low EPV or many uninfluential variables, or both. FPM and FLM had lower MSE than FM, as we would expect. However, the differences were relatively minor, especially when the EPV was high.

The values for the correctly specified model fitted to the breast cancer data suggest that there is potential to greatly improve the performance of a model by correctly specifying all functional relationships involving continuous predictors. This was apparent in the low EPV scenario where CSM achieved an agree value of 0.78 compared to the value of 0.62 achieved by FLM. In many scenarios, especially when EPV is low, it is difficult to model continuous predictors correctly. Prior medical knowledge can prove to be useful in these situations [4].

The performance of FM($R^2$), BE($\alpha$), and FPM($R^2$) depended on the scenario. If all the candidate predictors were influential then simplifying the model always reduced its prognostic ability. However, stepdown at the 0.99 level could still be safely employed in all scenarios to eliminate a few predictors with little change in the prognostic ability of the model. BE at a high significance level, for example 0.30, performed similarly. However, in the presence
of some uninfluential or noise predictors it was better to use lower stepdown and BE levels. Stepdown at the 0.95 or 0.90 level and BE at the 0.157 or 0.05 level often produced simplified models that performed better than the full model on our chosen criteria due to the removal of many uninfluential or weak predictors. FLM($R^2$) rarely performed better than FLM since uninfluential and weak predictors were often removed when the full model was fitted.

Choosing a simplified model based on the AIC criterion proved a successful strategy. In some situations the resulting model performed better on average, with respect to agreement and MSE, than the corresponding full model (for example, Figure 2), although they performed slightly worse with respect to concordance and the Brier score. However, the concordance and Brier scores of the full models were inflated compared to the correctly specified models. Simplification of the full model to a level determined by AIC resulted in concordance and Brier scores that tended to those of the CSM. We view this as a desirable consequence of simplification. Generally the results support the use of AIC for producing simplified models irrespective of the estimation and simplification methods used. Interestingly, simplification by BE gave almost identical results to simplification by stepdown. The advantage of stepdown appears to be its ease of application to prognostic indices from FPM and FLM.

Figure 6. Mean (a) agreement (b) MSE (c) concordance and (d) Brier score for FM ($R^2$), FPM ($R^2$), FLM ($R^2$) and BE ($\alpha$) at $R^2$ levels \{0.99, 0.95, 0.90, 0.80, 0.70\} and $\alpha$ levels \{0.50, 0.30, 0.157, 0.05, 0.001\} (the values are indicated from right to left on each line by $\triangle$). The full models are indicated by $\square$ on the relevant lines. Models selected by the automatic (AIC) strategy are indicated by $\bigcirc$. Simulated aortic aneurysm data scenario 3 ($n = 672$, EPV = 10, no noise variables).
In practice backwards elimination is often used with the significance level 0.05. In our simulations we found that this level was particularly suitable for the breast cancer data scenarios, where half the predictors were uninformative, and the aortic aneurysm scenarios involving noise variables, where almost half the predictors were uninformative. However the 0.05 significance level was not suitable for the aortic aneurysm scenarios without noise variables. Not surprisingly, higher significance levels that tended not to remove predictors performed far better in these cases. This might suggest that application of backwards elimination at the 0.05 level is inappropriate when we know from prior knowledge that all the candidate predictors are influential to some degree.

Similar work in this area has been performed by Steyerberg et al. [16]. They split a large data set (n = 40,830) into a training and test data set, and considered the performance of full models estimated using ML, penalized ML and the lasso. In addition they investigated BE at various levels, as well as a couple of other strategies, but did not consider any other simplification technique. They derived models in small subsets of the training data and assessed their performance, using c, model $\chi^2$ and calibration slope [23], in the test data set. This differs from our approach as we primarily assessed our models using a known truth. They concluded that stepwise selection has limited value and that shrinkage may improve performance substantially in small data sets. We agree with the latter but note that the predictors...
considered in their study were all reasonably prognostic and so that BE was not given the chance to remove weak or noise variables. In a survival analysis setting, Harrell et al. have also reported the poor performance of stepwise variable selection, assessed by concordance in validation data [24]. However, the limited investigation was in very low EPV scenarios (all <10) and a fixed \( z \)-level of 0.05 was chosen for the stepwise algorithm.

An obvious drawback of our work is that it is based on six scenarios derived from two data sets and hence may not be valid for other data sets. However the scenarios are different from one another yet the results produced are broadly consistent. In all cases simplification of the full model could reduce the number of predictors required with little loss or perhaps a gain in prognostic performance. In the presence of weak or noise predictors the savings in terms of number of variables were considerable.

This paper concentrated solely on predictive accuracy and did not consider statistical inference, for example confidence interval coverage and estimation of effects for individual variables. If inference were the focus of a study then we would suggest using a prespecified model. The advantage of using a prespecified model is that the regression coefficients are unbiased, the \( P \)-values are valid and the confidence intervals have the desired coverage. Harrell gives many other reasons for avoiding liberal use of variable selection [2].

All the simulations were performed in the software package Stata [25]. The BE strategy used built-in routines whereas the penalized and lasso simulations used specially written routines. These will published in the *Stata Journal* in due course but are currently available from the first author’s website (http://www.homepages.ucl.ac.uk/~ucakgam).

**APPENDIX**

Consider the linear predictor \( \hat{\eta} = \hat{\beta}_0 + \sum_j \hat{\beta}_j X_j \) where the predictors \( (X_j) \) are uncorrelated and have zero mean and unit variance. Regressing \( \hat{\eta} \) on the predictors would produce a perfect fit with \( R^2 = 1 \). Now the total sum of squares (about the mean) of \( \hat{\eta} \) is

\[
\text{TSS} = \sum_{i=1}^{n} \sum_{j=1}^{p} (\hat{\beta}_j x_{ij})^2
= \sum_{j=1}^{p} \hat{\beta}_j^2 \sum_{i=1}^{n} x_{ij}^2
\]

where \( x_{ij} \) is the value of \( X_j \) for the \( i \)-th individual, \( p \) is the number of predictors and \( n \) is the sample size.

The sample mean \( (\bar{x}_j) \) and variance \( (s_{x_j}^2) \) of each \( X_j \) are 0 and 1, respectively, so the expression for the sample variance

\[
s_{x_j}^2 = \frac{\sum_i x_{ij}^2 - n \bar{x}_j^2}{n-1}
\]

may be rearranged to give

\[
\sum_{i=1}^{n} x_{ij}^2 = n - 1
\]
Therefore

\[ \text{TSS} = (n - 1) \sum_{j} \hat{\beta}_j^2 \]

Omission of single predictor \( X_k \) would raise the residual sum of squares (RSS) from 0 to \( (n - 1)\hat{\beta}_k^2 \) so the decrease in \( R^2 \) would be

\[ \Delta R^2_k = \frac{\hat{\beta}_k^2}{\sum_j \hat{\beta}_j^2} \]

ACKNOWLEDGEMENTS

The first author received financial support from project grant number 045512/Z/95/Z from the Wellcome Trust. The second author was supported by the BUPA Foundation. We thank the editor and two referees for their comments which significantly improved the quality of this paper.

REFERENCES