STATISTICS IN MEDICINE
*Statist. Med.* 2000; **19**:1059–1079

# Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets

Ewout W. Steyerberg[1,*,†], Marinus J.C. Eijkemans[1], Frank E. Harrell Jr[2] and J. Dik F. Habbema[1]

[1] *Center for Clinical Decision Sciences, Department of Public Health, Erasmus University, Rotterdam, The Netherlands*
[2] *Division of Biostatistics and Epidemiology, Department of Health Evaluation Sciences, University of Virginia, Charlottesville, VA, U.S.A.*

## SUMMARY

Logistic regression analysis may well be used to develop a prognostic model for a dichotomous outcome. Especially when limited data are available, it is difficult to determine an appropriate selection of covariables for inclusion in such models. Also, predictions may be improved by applying some sort of shrinkage in the estimation of regression coefficients. In this study we compare the performance of several selection and shrinkage methods in small data sets of patients with acute myocardial infarction, where we aim to predict 30-day mortality. Selection methods included backward stepwise selection with significance levels $\alpha$ of 0.01, 0.05, 0.157 (the AIC criterion) or 0.50, and the use of qualitative external information on the sign of regression coefficients in the model. Estimation methods included standard maximum likelihood, the use of a linear shrinkage factor, penalized maximum likelihood, the Lasso, or quantitative external information on univariable regression coefficients. We found that stepwise selection with a low $\alpha$ (for example, 0.05) led to a relatively poor model performance, when evaluated on independent data. Substantially better performance was obtained with full models with a limited number of important predictors, where regression coefficients were reduced with any of the shrinkage methods. Incorporation of external information for selection and estimation improved the stability and quality of the prognostic models. We therefore recommend shrinkage methods in full models including prespecified predictors and incorporation of external information, when prognostic models are constructed in small data sets.


## 1. INTRODUCTION

Predictions from prognostic models may be used for a variety of reasons in medicine, including diagnostic and therapeutic decision making, selection of patients for randomized clinical trials,

* Correspondence to: Ewout W. Steyerberg, Center for Clinical Decision Sciences, Ee2091, Department of Public Health, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
† E-mail: steyerberg@mgz.fgg.eur.nl


*Received June 1998*
*Accepted July 1999*

and informing patients and their families [1]. The probability of a dichotomous outcome may well be estimated with a logistic regression model [2]. It is, however, often difficult to select predictors for such a prognostic model and estimate the regression coefficients for selected predictors correctly, that is, without overestimation [3]. These issues are especially prominent in relatively small data sets.

In this study, we aim to compare several methods for selection of predictors and estimation of logistic regression coefficients in small data sets. We distinguish selection and estimation methods which use only information from the data set under study (for example, stepwise selection, shrinkage), and methods which explicitly consider external information. We focus on predictive performance. Models are constructed in small parts of a large data set of patients with acute myocardial infarction, where we predict 30-day mortality with logistic regression analysis. We evaluate the models in an independent part of this data set.

We first describe the selection and estimation methods that we consider in this study (Section 2). The patient data are described in Section 3. Evaluations of predictive performance in data sets with around 23 or 62 events are presented in Sections 4 and 5. We further compare the performance of models with different numbers of covariables in Section 6. We discuss our findings in Section 7.

## 2. SELECTION AND ESTIMATION METHODS

We consider the usual logistic regression model log odds $\{Y = 1|X\} = \beta_0 + \Sigma\beta_i X_i$, where $Y$ is a binary outcome variable (0 or 1), $\beta_0$ is an intercept, and $\beta_i$ denotes the logistic regression coefficients for the design matrix $X$ of $i$ covariables. Our aim is to estimate the log odds $\{Y = 1|X\}$ accurately; interpretation of $\beta_i$ is secondary in our analyses. Table I gives an overview of the selection methods for the covariables in the matrix **X** and estimation methods for the regression coefficients $\beta$ that we consider in this study. We discuss the methods below.

### *2.1. Selection of predictors*

A large number of potentially prognostic covariables is often available in a prediction problem. Some of these may have been reported in the medical literature, some may be plausible predictors because of pathophysiologic mechanisms, others may simply be of interest to the investigator. Selection of a limited number of predictors is an obvious step. Data reduction is in concordance with the general scientific principle of parsimony, which implies that simpler models are more plausible descriptions of reality than more complex ones. Also, smaller models may be applied more easily in clinical practice [4].

Currently, stepwise selection methods are probably the most widely used in medical applications [5]. Forward selection starts with inclusion of the most significant candidate covariable, while backward selection starts with elimination of the least significant one from a full model. Forward and backward selection may also be combined. The stopping rule for inclusion or exclusion usually applies the standard significance level for testing of hypotheses ($\alpha = 0.05$), but the Akaike Information Criterion (AIC) has also been used [6]. Another class of stopping rules involves pooling of degrees of freedom of unselected predictors using the AIC or a residual chi-square test. An extension of the stepwise selection strategies is 'all possible subsets regression', where every possible combination of predictors is examined to find a best fitting model [7].

 

Table I. Overview of selection and estimation methods considered in this study.

| Method | Characteristics |
|---|---|
| *Selection of predictors* | |
| Full model | Inclusion of all candidate predictors, independent of statistical significance. |
| Backward stepwise | Selection based on the statistical significance of covariables in the data set under study. Leads to selection bias, since covariables with (by chance) large coefficients are more likely selected than those with (by chance) small coefficients. |
| Sign OK | Selection based on qualitative external information (plausibility of sign of the multivariable regression coefficient). |
| *Estimation of coefficients* | |
| Standard | Maximum likelihood estimation, leads to overestimation of regression coefficients for predictive purposes (estimation bias). |
| Shrunk | Linear shrinkage factor determined by bootstrapping. Bootstrapping corrects for selection bias by applying the selection procedure in every bootstrap sample. |
| Penalized | Penalty factor determined with Akaike Information Criterion (AIC) and effective degrees of freedom. Penalty factor for full model used in models constructed with a selection procedure. |
| Adapted | Uses quantitative external information (univariable regression coefficients from literature data) to adapt the standard coefficients |
| *Selection and estimation* | |
| Lasso | Selection because some coefficients are shrunk to zero. |

Stepwise methods identify a limited number of covariables for inclusion in regression models, which may be considered as the most important predictors in a prognostic problem. However, several drawbacks are known for these techniques [8, 9]. The selection is unstable, in the sense that the addition or deletion of a small number of patients can alter it, especially when covariables are correlated [10–13]. Further, stepwise selection has limited power to select prognostically important covariables in small data sets, which will lead to a loss in predictive ability. On the other hand, there is a substantial risk that one or more (almost) random covariables are selected, since multiple comparisons are made [8, 9]. Next, the variance of the coefficients is usually calculated as if the selection was predetermined. This causes underestimation of standard errors and $p$-values in the resulting model [12, 14, 15].

Given the problems with stepwise selection methods, alternatives have been considered. The most obvious selection strategy is to fit a fixed selection of predefined predictors, for example, based on firm clinical knowledge and information from other studies [2]. We refer to this approach as fitting a 'full model' (Table I). We include continuous variables as linear terms, assuming linearity on the log-odds scale. Fitting main effects further assumes additivity of the predictors. For simplicity, these assumptions were not assessed in our evaluation, as would be possible with the extension of the models with non-linear and interaction terms [1]. As an intermediate between fitting a full model and stepwise selection with the standard significance level ($\alpha = 0.05$), we may apply stepwise selection with a high $\alpha$ for selection. We may for example exclude covariables with $p$-values exceeding 0.50, arguing that these probably contribute more

 

noise than predictive information to the model. In our evaluation, we apply stepwise selection with backward elimination of predictors from a full model, using $\alpha = 0.50$, 0.157, 0.05 and 0.01. The $\alpha$ of 0.157 corresponds to the use of the AIC, since all covariables had 1 degree of freedom.

### *2.2. Estimation of regression coefficients in full models*

A key problem of regression modelling in small data sets is that the regression coefficients are overestimated for predictive purposes. When the standard maximum likelihood estimates of the logistic regression coefficients are shrunk towards zero, predictions will show a better calibration in new patients [3]. A simple and somewhat crude approach is to apply a linear shrinkage factor for the regression coefficients. The shrinkage factor may be based on a heuristic formula [1, 3, 16]; [model $\chi^2$ − (d.f. −1)]/model $\chi^2$, where d.f. indicates the degrees of freedom of the covariables fit in the model. The required shrinkage increases when larger numbers of predictors are considered (d.f. ↑), or when the sample size is smaller (model $\chi^2$ ↓) [3, 16].

We calculated a linear shrinkage factor for the regression coefficients with bootstrapping [1, 17, 18].

1. Take a random bootstrap sample of the same size as the original sample, drawn with replacement.
2. Select the covariables according to the selection procedure and estimate the logistic regression coefficients in the bootstrap sample.
3. Calculate the value of the prognostic index for each patient in the original sample. The prognostic index is the linear combination of the regression coefficients as estimated in the bootstrap sample with the values of the covariables in the original sample. The prognostic index indicates the expected log odds of the outcome in the original sample, and is equivalent to the 'linear predictor' in the context of general linear models.
4. Estimate the slope of the prognostic index with logistic regression, using the outcomes of the patients in the original sample.

Steps 1 to 4 were repeated 300 times to obtain a stable estimate of the shrinkage factor, which was calculated as the mean of the 300 slopes estimated in step 4. 'Shrunk' coefficients were calculated by multiplication of the standard coefficients with the shrinkage factor, which might take values between 0 and 1.

Further, shrinkage may be achieved by inclusion of a penalty factor $\lambda$ in the maximum likelihood formula [19–21]: $\log L - \frac{1}{2}\lambda\beta' P\beta$. Here $L$ denotes the usual likelihood function, $\lambda$ is the (positive) penalty factor, $\beta'$ denotes the transpose of the vector of estimated regression coefficients $\beta$ (excluding the intercept), and $P$ is a penalty matrix. In our analyses, the diagonal of $P$ consisted of the variances of the covariables and all other values of $P$ were set to zero. This choice of $P$ makes the penalty to the log-likelihood unitless. This scaling was used both for continuous and dichotomous covariables, although dichotomous variables might generally not require scaling by their variance. To determine the optimal value of $\lambda$, we varied $\lambda$ over a grid, for example, 0, 0.5, 1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 32, and evaluated a modified AIC: [model $\chi^2 - 2 \times$ effective d.f.]. A reviewer noted that penalized ML is very similar to applying a linear shrinkage factor $c$ when the matrix $P$ is equal to the full matrix of second derivatives, with $c = 1/(1 + \lambda)$. Details of the penalized ML procedure were described before [1, 28, 33].

 

### 2.3. Estimation of regression coefficients after selection

Stepwise selection methods cause predictors with relatively large regression coefficients to be more likely selected than predictors with relatively small regression coefficients. The process of estimation after testing ('testimation' [22]), leads to overestimation of the regression coefficients of predictors included in the final model [5, 9, 23]. This selection bias should be taken into account when calculating a shrinkage factor. This may be achieved by considering the number of candidate predictors in the heuristic formula (instead of the number of selected predictors) [3]. For our evaluations, we included the selection process in step 2 of the bootstrapping procedure as described above [1]. For penalized estimates of the regression coefficients after selection, we applied the penalty factor that was identified as optimal for the full model.

Further, techniques have recently been developed which select predictors by shrinking some coefficients to zero [24–26]. We applied the 'Lasso' (least absolute shrinkage and selection operator), which can readily be applied to linear regression models but also to generalized linear models such as the logistic or Cox model [24, 25]. The Lasso estimates the regression coefficients $\beta$ of standardized covariables by minimizing the log-likelihood subject to $\Sigma|\beta| \leqslant t$, where $t$ determines the shrinkage in the model. We varied $s = t/\Sigma|\beta^0|$ over a grid from 0.5 to 0.95, where $\beta^0$ indicates the standard ML regression coefficients and $s$ may be interpreted as a standardized shrinkage factor. We estimated $\beta$ with the value of $t$ that gave the lowest mean-squared error in a generalized cross-validation procedure [24].

### 2.4. External information

For selection and estimation,we may use information from outside the data set under study, for example, quantitative results from published studies. Hereto, we previously developed an 'adaptation' method, which enabled us to combine estimates of the univariable regression coefficients from published studies with multivariable regression coefficients estimated in the data set under study [27, 28]. The formula for the adapted coefficients is

$$\beta_{\mathrm{m}|\mathrm{L}} = \beta_{\mathrm{m}|\mathrm{I}} + c(\beta_{\mathrm{u}|\mathrm{L}} - \beta_{\mathrm{u}|\mathrm{I}})$$

where $\beta_{\mathrm{m}|\mathrm{L}}$ indicates the multivariable coefficients, adapted for univariable literature information, $\beta_{\mathrm{m}|\mathrm{I}}$ indicates the multivariable coefficient estimated in the data set with individual patient data (standard ML estimates), and $\beta_{\mathrm{u}|\mathrm{L}}$ and $\beta_{\mathrm{u}|\mathrm{I}}$ indicate the univariable coefficients in the literature data and individual patient data. The adaptation factor '$c$' is estimated as

$$c = \rho(\mathrm{uni}, \mathrm{mult})\,[\mathrm{SE}(\beta_{\mathrm{m}|\mathrm{I}})\mathrm{SE}(\beta_{\mathrm{u}|\mathrm{I}})]/[\mathrm{var}(\beta_{\mathrm{u}|\mathrm{L}}) + \mathrm{var}(\beta_{\mathrm{u}|\mathrm{I}})]$$

where $\rho(\mathrm{uni}, \mathrm{mult})$ indicates the correlation between univariable and multivariable regression coefficients (estimated from 300 bootstrap samples), SE indicates the estimated standard error, and var the estimated variance. The value of '$c$' can also simply be set to 1 [29], which gave very similar results in our analyses [28].

Further, we might make assumptions on the direction of the effect of a predictor. This may be difficult for the multivariable context, since correlations between predictors may explain a counterintuitive sign of a regression coefficient. However, for many prognostic problems it may be reasonable to suppose that predictors that indicate a higher risk in univariable analyses also increase a patient's risk while adjusting for other predictors in a multivariable analysis. We might

 

therefore select only predictors in the multivariable model which have a multivariable coefficient with a sign that is identical to the sign in univariable analyses in the literature. We realize that this may introduce some selection bias, since chance dictates that coefficients sometimes have a counterintuitive sign. Exclusion of these predictors hence gives some bias to higher values of regression coefficients of selected predictors.

We note that this 'Sign OK' approach has some similarity with what has been labelled 'Bayes-empirical-Bayes estimation [30]. It may also prevent what epidemiologists have labelled the type III error, that is, the inclusion of covariables with an incorrect sign [31].

We applied 'Sign OK' selection on full models, and on models containing predictors with $p$-values $< 0.50$ as selected with backward stepwise selection. When the multivariable sign was different from the univariable sign, the predictor was excluded from the model. The model was refitted, and the sign of the remaining covariables checked. This procedure was repeated until all selected predictors had a correct sign.

### *2.5. Small data sets*

For testing of logistic regression coefficients, the statistical power is predominantly determined by the smallest of the two frequencies of the binary outcome. To prevent problems of overfitting, or more specifically, overestimation of regression coefficients, the number of candidate covariables considered should be in reasonable balance with the number of events [1, 10]. It has been suggested as a general rule, but without full study, that the number of events per variable (EPV) should be at least 10 [4, 10, 31, 32]. In this study we focus on small data sets, which means that logistic models are constructed in data sets where the 1:10 rule is violated. For comparison, we also study models in the situation that the EPV exceeds 10, 20 or 50.

### *2.6. Software*

All calculations were performed with S-plus software (version 3.3, MathSoft, Inc., Seattle WA) and/or SAS 6.08 (SAS Institute Inc., Cary NC). We used Harrell's Design [33, 34] and Tibshirani's Lasso library [35]. An example of a S-plus program including the various modelling approaches is available in the public domain [36].

## 3. EMPIRICAL EVALUATION

### *3.1. Patients*

For evaluation of the selection and estimation methods we used the data from 40 830 patients with complete follow-up from the GUSTO-I clinical trial [37, 38]. In brief, this data set consists of patients with an acute myocardial infarction, who were randomized to one of four thrombolytic regimens. The differences between these regimens were small relative to the effect of predictive covariables, and are ignored in the present analyses. Mortality at 30 days was the primary endpoint, and occurred in 2851 patients (7.0 per cent). Within the total data set, we distinguished 16 regions: eight in the United States (U.S.); six in Europe, and two others (Canada and Australia/New Zealand). These regions included 2552 patients and 178 deaths on average. Within regions, 'large' and 'small' multi-centre subsamples were created by grouping hospitals together

 

on a geographical basis. The large subsamples were created such that they each contained at least 50 events. The subgrouping procedure was repeated to create small subsamples with at least 20 events. These subsamples are not strictly random samples, but aimed to reflect the real-life situation where a small multi-centre data set containing patients from several nearby hospitals is available to construct a prognostic model which should be applicable to the total patient population.

The data set was split into a training and test part [28]. These parts each consisted of eight regions with geographical balance and similar overall mortality (7.0 per cent). The training part ($n = 20\,512$) contained 61 small and 23 large subsamples containing on average 336 and 892 patients of whom 23 and 62 died, respectively. Logistic regression models were constructed in the subsamples from the training part and evaluated in the test part. We do not advise analysts to hold back data for model validation in routine practice, as internal validity can more efficiently be studied with re-sampling techniques such as the bootstrap [17, 18, 39]. For our evaluation, an independent test part was, however, considered most convenient.

### 3.2. *Predictors considered*

We considered previously defined prognostic models for acute MI. The selection of predictors was hence kept external to the findings in the GUSTO-I data set. We focused on evaluations of an eight-predictor model, as defined in the TIMI-II study [40]. This included shock, age $> 65$ years, anterior infarct location, diabetes, hypotension, tachycardia, no relief of chest pain, and female gender. The dichotomization of predictors is generally not advisable, as will be illustrated empirically (Section 6). A three-predictor model was considered that contained the continuous variables age, Killip class (a measure for left ventricular function) and the dichotomous variable anterior infarct location [41]. In the GISSI-2 data set ($n = 9720$; 772 in-hospital deaths), the continuous variable 'number of leads with ST elevation' was selected in addition to these three variables [42]. In the large subsamples and in the regions, we further considered a 17-predictor model, consisting of the TIMI-II model plus nine other covariables considered in previous analyses [38, 40–43].

Table II shows the distribution of the 17 covariables and their uni-variable and multivariable logistic regression coefficients in the eight-predictor and 17-predictor model. Results are shown for the training part only, since results for the test part were very similar. We note that the dichotomous predictors 'hypotension' and 'shock' had a low prevalence, but a strong effect in the multivariable models. This low prevalence led to zero cells and non-convergence of the eight-predictor model in 12 of the 61 small subsamples. These subsamples were excluded from the evaluations. Most multivariable coefficients were smaller (closer to zero) than the univariable coefficients, reflecting (modest) positive correlations between the predictors ($r$ generally around 0.1–0.2). All predictors had an identical sign in univariable and multivariable analyses. The signs shown in Table II were used for 'Sign OK' selection. Note that some predictors had a sign that might not have seemed plausible *a priori*; smokers, patients with hypercholesterolaemia or a family history of MI were at a decreased risk of 30- day mortality. This findings confirms that risk factors for developing the disease (acute MI) do not have to correspond to prognostic factors in patients with the disease. All coefficients in the eight-predictor model were significant at $p < 0.001$. The 17-predictor model contained covariables with relatively small coefficients, such as sex, hypertension, previous angina and family history in the training part.

 

Table II. Distribution of predictors and logistic regression coefficients (standard error) in the eight-predictor and 17-predictor models. Results are shown for the training sample ($n = 20\,512$, 1423 died) in the GUSTO-I data set.

| Predictors | Prevalence* | Logistic regression coefficients | | |
|---|---|---|---|---|
| | | Univariable | 8-predictor model | 17-predictor model |
| Age > 65 years | 41% | 1.52 (0.06) | 1.38 (0.06) | 1.14 (0.07) |
| Female sex | 24% | 0.77 (0.06) | 0.45 (0.06) | 0.08 (0.09) |
| Diabetes | 13% | 0.58 (0.07) | 0.27 (0.08) | 0.29 (0.08) |
| Hypotension (BP < 100 mmHg) | 8% | 1.28 (0.07) | 1.24 (0.08) | 1.25 (0.08) |
| Tachycardia (pulse > 80) | 31% | 0.75 (0.06) | 0.67 (0.06) | 0.65 (0.06) |
| Anterior infarct location | 39% | 0.93 (0.06) | 0.76 (0.06) | 0.43 (0.07) |
| Shock (Killip III/IV) | 2% | 2.51 (0.10) | 1.74 (0.12) | 1.69 (0.12) |
| No relief of chest pain | 65% | 0.55 (0.06) | 0.52 (0.07) | 0.53 (0.07) |
| Previous MI | 16% | 0.79 (0.06) | | 0.59 (0.07) |
| Height ($\times 10$ cm)† | 17.1 | −0.47 (0.03) | | −0.16 (0.05) |
| Weight ($\times 10$ kg)† | 7.9 | −0.29 (0.02) | | −0.11 (0.03) |
| Hypertension in history | 37% | 0.30 (0.06) | | 0.11 (0.06) |
| Smoking†‡ | 1.9 | 0.48 (0.03) | | 0.17 (0.04) |
| Hypercholesterolaemia | 35% | −0.27 (0.06) | | −0.18 (0.07) |
| Previous angina | 37% | 0.40 (0.06) | | 0.14 (0.06) |
| Family history | 41% | −0.37 (0.06) | | −0.13 (0.06) |
| ST elevation in > 4 leads | 37% | 0.65 (0.06) | | 0.35 (0.07) |

* Percentage of patients with the characteristic or average value (continuous variables).
† Continuous predictor, modelled as linear term in logistic regression analysis.
‡ Smoking was coded as 1 for current smokers, 2 for ex-smokers, 3 for never smokers.

### *3.3. Evaluation*

The evaluation of model performance considered discrimination, calibration and overall performance [1, 44]. Discrimination refers to the ability to distinguish high risk patients from low risk patients, and is commonly quantified by a concordance statistic ($c$) [45]. In logistic regression $c$ is identical to the area under the receiver operating characteristic (ROC) curve.

Calibration refers to whether the predicted probabilities agree with the observed probabilities. Several 'goodness-of-fit' statistics are available to quantify calibration [3, 44, 46]. We used the slope of the prognostic index, since this measure is readily interpretable in the context of overestimation of regression coefficients. The prognostic index was calculated as the linear combination of the regression coefficients as estimated in the subsample with the values of the covariables in the test part. Models with overestimated regression coefficients will show a slope of the prognostic index which is less than 1, indicating that low predictions are too low, and high predictions are too high.

Finally, we quantified the overall model performance in one number, the model $\chi^2$, which is closely related to the Kullback–Leibler distance [47]. The model $\chi^2$ was calculated as the difference between the 2-log-likelihood of a model with an intercept and the prognostic index as an offset variable (slope fixed at unity, that is, the prognostic index was taken literally), and the −2-log-likelihood of a model with an intercept only. A negative model $\chi^2$ implied that a model performed worse than predicting the average risk for every patient.

 

Table III. Logistic regression coefficients and evaluation of performance for a small subsample (429 patients, 24 died, see text).

| | Full model | | | Backward stepwise, $\alpha = 0.05$ | | | Lasso | Full | Stepwise | Lasso |
|---|---|---|---|---|---|---|---|---|---|---|
| | Standard | Shrunk | Penalized | Standard | Shrunk | Penalized | | Gold standard | | |
| *Predictors* | | | | | | | | | | |
| Shock | 2.96 | 2.44 | 2.71 | 3.67 | 2.84 | 3.22 | 2.90 | 1.73 | 2.07 | 1.75 |
| Age > 65 years | 1.37 | 1.13 | 0.93 | 1.36 | 1.05 | 0.94 | 1.11 | 1.37 | 1.49 | 1.46 |
| Anterior MI | 0.76 | 0.62 | 0.55 | 0 | 0 | 0 | 0.58 | 0.76 | 0 | 0.73 |
| Diabetes | −0.11 | −0.09 | 0.03 | 0 | 0 | 0 | −0.03 | 0.29 | 0 | 0.32 |
| Hypotension | 1.39 | 1.15 | 1.04 | 1.19 | 0.92 | 0.94 | 1.16 | 1.25 | 1.12 | 1.26 |
| Tachycardia | 0.89 | 0.73 | 0.61 | 0 | 0 | 0 | 0.69 | 0.66 | 0 | 0.69 |
| No relief | 0.68 | 0.56 | 0.43 | 0 | 0 | 0 | 0.50 | 0.55 | 0 | 0.51 |
| Sex | −0.04 | −0.03 | 0.04 | 0 | 0 | 0 | 0 | 0.44 | 0 | 0 |
| *Model performance in test sample* ($n = 20\,318$) | | | | | | | | | | |
| Area under ROC | 0.77 | 0.77 | 0.78 | 0.74 | 0.74 | 0.73 | 0.78 | 0.79 | 0.74 | 0.79 |
| Slope | 0.75 | 0.91 | 0.91 | 0.71 | 0.92 | 0.83 | 0.84 | 0.94 | 0.97 | 0.95 |
| Model $\chi^2$ | 1279 | 1399 | 1341 | 941 | 1101 | 987 | 1323 | 1604 | 1250 | 1558 |

## 4. RESULTS IN SMALL DATA SETS

### 4.1. Illustration: a small subsample

We illustrate the use of the selection and estimation methods with a small subsample, which showed results that were typical for the other small subsamples [36]. Table III shows the regression coefficients as estimated in the subsample, and the performance in the test part, which was independent from the subsample ($n = 20\,318$). The subsample was created by combining the patient data from 10 hospitals in the Western region of the U.S.A. that participated in the GUSTO-I trial. The sample included 429 patients, of whom 24 died.

The full eight-predictor model had large estimated regression coefficients for shock, age and hypotension. The coefficients were shrunk with a factor 0.824 according to the bootstrapping procedure. Penalized estimates of the regression coefficients were obtained with a penalty factor of 8. Backward stepwise selection with $\alpha = 0.05$ led to the inclusion of three predictors. Shrinkage was calculated as 0.773, in a bootstrapping procedure which included the stepwise selection in each bootstrap sample. The Lasso parameter 's' was 0.8375, and resulted in the coefficient of 'sex' to be set at 0. Hence, seven predictors were selected, with coefficients that were shrunk compared to the standard estimates in the full model. As a reference, the final columns show the coefficients obtained in the total training part ('gold standard', $n = 20\,512$).

The performance of the stepwise models was worse than the full models, with respect to the area under the ROC curve and overall performance as indicated by the model $\chi^2$. As expected, the slope of the prognostic index was closer to 1 for the shrunk and penalized models than the standard ML estimates. The Lasso performed similarly to the shrunk or penalized full model. The performance of the gold standard models in the evaluation sample indicates

 

Table IV. Logistic regression coefficients and evaluation of model performance, combining data from a small subsample (429 patients, 24 died) with external information (see text).

| | Sign OK | | | Full | Stepwise | Sign OK |
|---|---|---|---|---|---|---|
| | Standard | Shrunk | Penalized | Adaptation | | |
| *Predictors* | | | | | | |
| Shock | 2.94 | 2.42 | 2.71 | 2.20 | 2.89 | 2.18 |
| Age > 65 years | 1.36 | 1.12 | 0.93 | 1.75 | 1.71 | 1.74 |
| Anterior MI | 0.76 | 0.62 | 0.55 | 0.62 | 0 | 0.62 |
| Diabetes | 0 | 0 | 0.03 | −0.12 | 0 | 0 |
| Hypotension | 1.38 | 1.13 | 1.04 | 1.59 | 1.39 | 1.58 |
| Tachycardia | 0.87 | 0.72 | 0.61 | 0.63 | 0 | 0.61 |
| No relief | 0.67 | 0.55 | 0.43 | 0.59 | 0 | 0.59 |
| Sex | 0 | 0 | 0.04 | 0.59 | 0 | 0.63 |
| *Model performance in test sample* ($n = 20\,318$) | | | | | | |
| Area under ROC | 0.78 | 0.78 | 0.78 | 0.78 | 0.74 | 0.79 |
| Slope | 0.76 | 0.92 | 0.91 | 0.79 | 0.77 | 0.80 |
| Model $\chi^2$ | 1308 | 1422 | 1341 | 1481 | 1131 | 1496 |

that the stepwise model, which contained only three of the eight predictors, might have a model $\chi^2$ of 1250 if regression coefficients were (almost) correctly estimated. By contrast, the Lasso selection could achieve a model $\chi^2$ of 1558, which is close to the performance of the full eight-predictor model (1604). These findings are consistent with the omission of only one predictor (Lasso) or five predictors (stepwise) from the full model.

Table IV shows the results for the same subsample when qualitative or quantitative external information was taken into account. 'Sign OK' selection led to the exclusion of the predictors 'diabetes' and 'sex' from the standard or shrunk full model. This exclusion improved the model performance slightly. No implausible signs were noted for the penalized full model. For the adaptation method, univariable regression results were used from the training sample, excluding the subsample ($n = 20\,512 - 429 = 20\,083$ patients). The estimated correlation between univariable and multivariable regression coefficients was around 0.85 for most covariables (range 0.80–0.98), and the adaptation factors were between 0.95 and 1.0. For most predictors in the full or stepwise models, the adapted regression coefficients were somewhat closer to the gold standard coefficients. Model performance improved, with the best model $\chi^2$ with sign OK selection and adapted estimation of the regression coefficients. We note that the adaptation method may show an unrealistically favourable performance in our evaluations, since the comparability between the subsample and the 'literature data' will be higher than generally may be expected in practice. The results from the analyses with adapted coefficients should be considered as indicating the maximal obtainable benefit by this estimation method.

### *4.2. Numbers of covariables selected*

In the small subsamples, backward stepwise selection on average led to the inclusion of 1.8 predictors with $\alpha = 0.01$, and to selection of 5.2 predictors with $\alpha = 0.50$ (Table V). When we

 

Table V. Number of predictors selected from 8 or 17 candidate predictors (averages (standard deviation)). Results shown for standard ML estimation in 61 small subsamples (336 patients, 23 deaths on average) and 23 large subsamples (892 patients, 62 deaths on average).

| Predictors | Sample size | Backward stepwise | | | | Full | Sign OK | | Lasso |
|---|---|---|---|---|---|---|---|---|---|
| | | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.157$ | $\alpha = 0.50$ | | $\alpha = 0.50$ | Full | |
| 8 | small | 1.8 (0.9) | 2.3 (1.0) | 2.7 (1.2) | 5.2 (1.9) | 8 | 5.1 (1.8) | 7.0 (0.9) | 7.5 (0.6) |
| 8 | large | 4.2 (1.0) | 5.0 (1.0) | 5.7 (1.3) | 7.0 (1.2) | 8 | 7.0 (1.2) | 7.8 (0.4) | 7.9 (0.4) |
| 17 | large | 4.7 (1.5) | 5.9 (1.4) | 6.8 (1.7) | 8.9 (2.1) | 17 | 7.7 (1.8) | 14.3 (1.3) | 16.3 (1.0) |

required that the sign of the coefficient had to be correct ('Sign OK'), the selection with $\alpha = 0.50$ remained similar. From the full model, 'Sign OK' selection on average excluded 1.0 predictor with standard or shrunk estimation, and 0.8 or 0.3 predictors with penalized or adapted estimation, respectively. The Lasso led to the selection of 7.5 predictors, with an average value of 0.80 for the Lasso parameter 's'.

### *4.3. Average performance*

Figure 1 shows the average performance of the models constructed in the small subsamples when evaluated in the test part. In addition to the results shown in Tables III and IV, we show the performance of backward stepwise selection with $\alpha = 0.01$, 0.157 or 0.50. A full model may be interpreted as stepwise selection with $\alpha = 1.0$. For the stepwise and 'Sign OK' models, we show results with standard, shrunk, penalized and adapted regression coefficients.

The area under the ROC curve ($c$) increased when a higher $\alpha$ was used for stepwise selection (range 0.01 to 0.50). Areas for the models with standard or shrunk coefficients are by definition identical, since the ordering of the predicted probabilities does not change by applying a linear shrinkage factor. The penalized models performed similarly to the standard or shrunk models, except for the full model, where penalized models were somewhat better (area 0.760 versus 0.754). Selection of covariables with the correct sign improved the performance for the full models. The Lasso performed similar to the selection of a full model with shrunk or penalized estimation of regression coefficients. Adapted models performed best for all selection strategies. If covariables were selected with a correct sign from full models ('Sign OK, full'), the average $c$ (0.782) was very close to the performance of the gold standard model ($c = 0.786$).

The slope of the prognostic index should ideally be 1.0; it was 0.944 for the gold standard model, consistent with the observation that the regression coefficients were somewhat larger in the training part than in the test part. The standard ML estimates had slopes around 0.7. This figure was the result of the combination of overestimation for predictive purposes (in a prespecified model) and selection bias (caused by stepwise selection). Apparently, when a higher $\alpha$ was used for selection, that is, going from 0.01 to 0.50, the decrease in selection bias was offset by an increase in overestimation bias, such that the sum of both biases was approximately constant. The shrunk or penalized estimates were better calibrated, although some overestimation remained when stepwise selection was applied with a low $\alpha$. The adapted estimates showed a remarkable pattern. Selection with a low $\alpha$ led to a better calibration than selection with $\alpha = 0.50$ or a full model. This may be explained by the extent of estimation bias, which was very limited when only

 

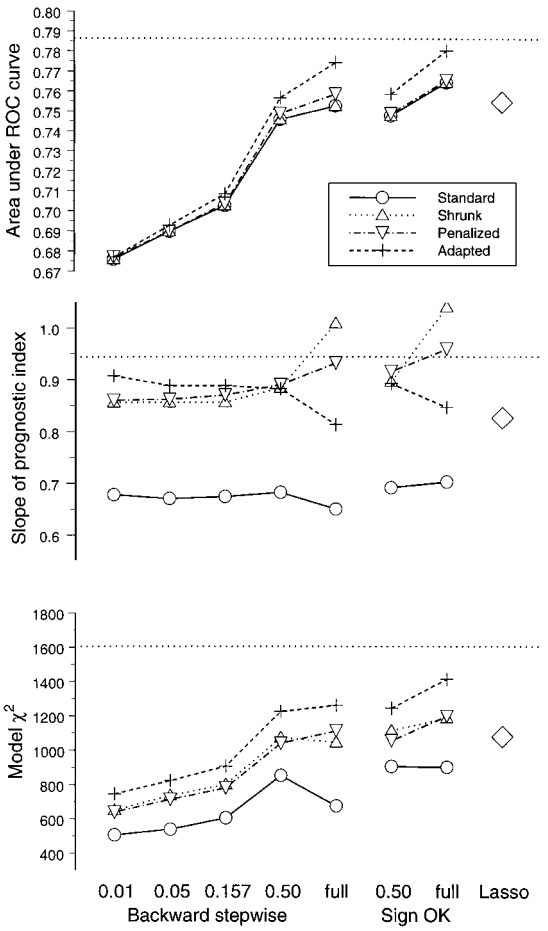


Figure 1. Average performance of models constructed in small subsamples (average: 23 deaths) of the GUSTO-I data set. Eight dichotomous predictors were considered. For stepwise and 'Sign OK' selection, results are shown for standard, shrunk, penalized and adapted estimation of the logistic regression coefficients (see text). The dotted line indicates the performance in the test part for the model estimated in the total training set and serves as a reference.

a few covariables were selected (for example, $\alpha = 0.01$), and was largest when full models were fit. Some shrinkage might hence improve the performance of the adapted models [28]. The Lasso led to an average slope of 0.83.

Finally, we consider the overall model performance, as expressed by the model $\chi^2$ in the test data set. The gold standard model had a model $\chi^2$ of 1604. The patterns in overall model performance closely reflect the patterns observed with respect to the area under the ROC curve. The main difference is the poor performance of the models with standard ML estimates, especially full models. This is explained by their poor calibration. The average Lasso model $\chi^2$ was 1079, compared to 1073 for $\alpha = 0.50$ selection and shrunk estimation, and 1112 for penalized full models.

 

## 5. RESULTS IN LARGER DATA SETS

### 5.1. Numbers of covariables selected

Starting with the eight-predictor model, backward stepwise selection on average led to the inclusion of 4.2 to 7.0 predictors with $\alpha = 0.01$ to 0.50 in the large subsamples (Table V). The Lasso led to the selection of 7.9 predictors on average, with an average value of 0.87 for the Lasso parameter 's'.

More predictors were selected when the 17-predictor model was considered. The variability increased too, as indicated by the larger standard deviation. 'Sign OK' selection excluded around one predictor from those selected with $\alpha = 0.50$ (7.7 instead of 8.9), and around two from a full 17-predictor model. The Lasso parameter 's' was 0.81 on average, leading to an average selection of 16.3 of the 17 covariables.

### 5.2. Performance of eight-predictor models

Figure 2 shows the average performance of the 23 eight-predictor models constructed in the large subsamples. The area under the ROC curve increased when a higher $\alpha$ was used for stepwise selection and was highest for full models. The areas for the models with standard, shrunk or penalized coefficients were very similar. Selection of covariables with the correct sign had a negligible effect, which is explained by the fact that most covariables already had the correct sign in the full model and in $\alpha = 0.50$ selected models (see Table V). The Lasso had the highest ROC area of the methods that use only information from the data set under study. Models with adapted coefficients performed best for all selection strategies.

The slope of the prognostic index was clearly less than 1 for the models with standard estimates of the regression coefficients. The Lasso performed very well with an average slope of 1.01.

The model $\chi^2$ reflects the patterns seen with discrimination and calibration. For the estimation methods, the worst performance was seen with the standard ML estimates of the regression coefficients, and the best performance with the adaptation method. Selection of full models, or with the Lasso, was much better than backward selection with a low $\alpha$ (0.01/0.05).

When we compare Figure 2 with Figure 1, we note that the performance for all models increased, which is explained by the larger number of patients considered. Especially, backward stepwise methods performed much better in the large subsamples than in small subsamples. For example, the model $\chi^2$ increased from around 600 to over 1200 for the models with $\alpha = 0.05$ selection and standard estimates. In the large subsamples, full models performed better than stepwise selected models, while $\alpha = 0.50$ selection was better than a full model in the small subsamples. The relative improvement was smaller for the Lasso (model $\chi^2$ from 1079 to 1517) and for the 'Sign OK' selection in full models with adapted estimation of the regression coefficients (from 1397 to 1558). The latter methods achieved a model quality in the small subsamples which was only achieved by other methods in data sets that were over twice as large.

### 5.3. Performance of 17-predictor models

Figure 3 shows the average performance of the 23 17-predictor models constructed in the large subsamples. We focus on the comparison with Figure 2, where eight predictors were considered from the set of 17 in Figure 3. We would expect that consideration of more predictors would increase the performance of the model. This was not the case when backward stepwise selection

 

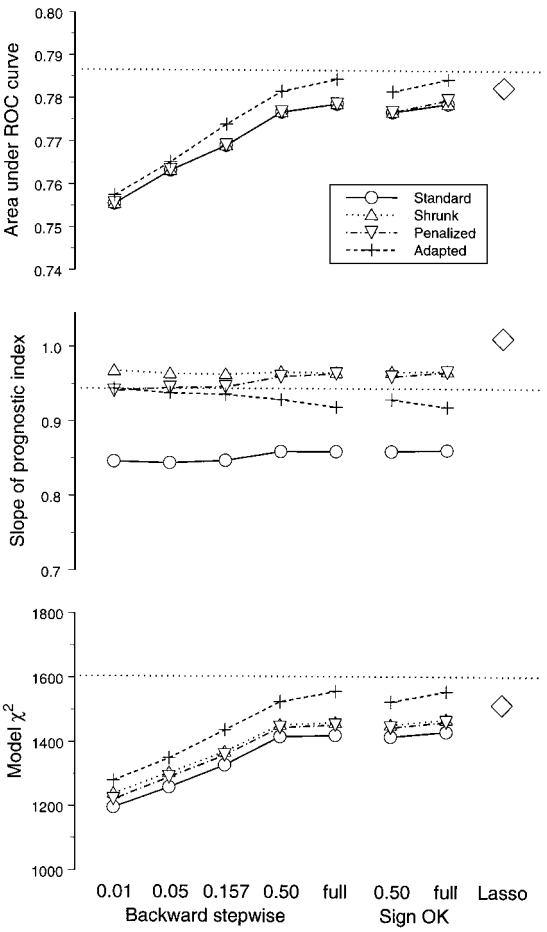


Figure 2. Average performance of models constructed in large subsamples (average: 62 deaths) of the GUSTO-I data set. Eight dichotomous predictors were considered. The dotted line indicates the performance in the test part for the model estimated in the total training set and serves as a reference.

was used. For $\alpha$ ranging from 0.01 to 0.50, the area under the ROC curve was around 0.01 lower for all estimation methods. The slope of the prognostic index remained around 1 for the shrunk or penalized models, but was further away from 1 for the standard or adapted models. The model $\chi^2$ decreased by around 150.

For the full models, a slightly worse performance was noted for the standard or shrunk estimates, and a slightly better performance for the penalized or adapted estimates. 'Sign OK' selection was helpful; when only predictors with a correct sign were included from the 17 covariables, model performance was better than selection from eight covariables with the correct sign. Finally, the Lasso performed slightly worse when starting with 17 instead of eight covariables (model $\chi^2$ decreased from 1517 to 1492).

 

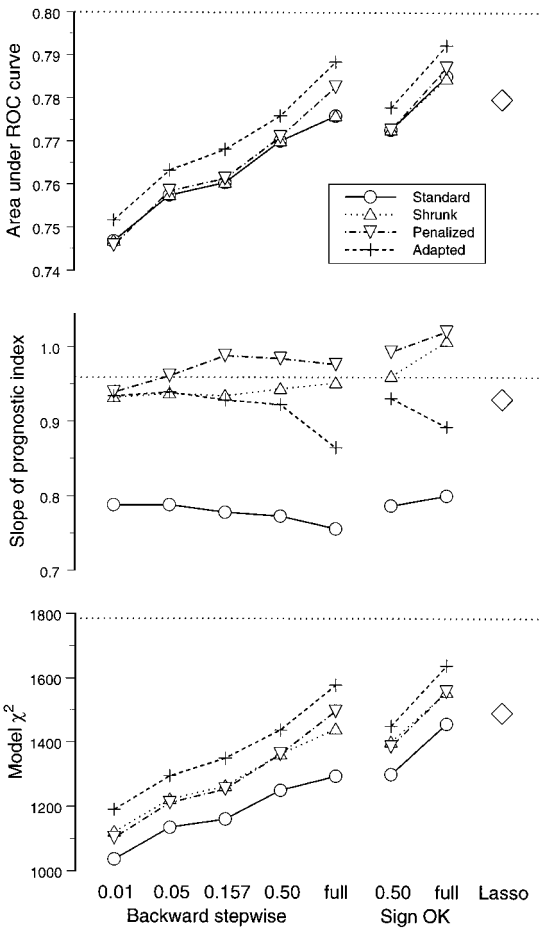


Figure 3. Average performance of models constructed in large subsamples (average: 62 deaths) of the GUSTO-I data set. 17 predictors were considered, including the eight dichotomous predictors from Figure 2. The dotted line indicates the performance in the test part for the model estimated in the total training set and serves as a reference.

### 5.4. *Larger subsamples: regions*

The eight- and 17-predictor models were also constructed with the selection and estimation methods in the eight regions distinguished in the test part of the GUSTO-I data set. Since the number of events was 178 on average, these models fulfilled the 1:10 criterion and no major problems might be expected. Indeed, differences between the different selection and estimation methods were small. However, the average performance of full 17-predictor models was better than stepwise selected models (starting with the 17-predictor model). For example, the average model $\chi^2$ was 1634 for a full 17-predictor model, and 1546 when stepwise selection with $\alpha = 0.05$ was applied (both with standard ML estimates). Shrunk, penalized, adapted or Lasso estimates were only slightly better than the standard ML estimates. The full 17-predictor model performed

 

better than the full eight-predictor model. For example, the average model $\chi^2$ was 1660 and 1548 for the full 17- and eight-predictor model with shrunk estimation of the regression coefficients, and 1675 and 1545 with penalized estimation, respectively. In the full eight-predictor model, use of the standard ML estimates of the regression coefficients led to an average performance (model $\chi^2$ 1537) that was close to shrunk or penalized estimation (model $\chi^2$ 1548 or 1545, respectively). Apparently, when the number of events per variable exceeded 20 ($178/8 \approx 22$), the standard ML estimates in a full model were satisfactory.

## 6. COMPARISON OF DIFFERENT FULL MODELS

Comparison of Figure 3 with Figure 2 indicated that considering more predictive covariables did not imply that model performance improved. To provide more insight, we compared full models including 3, 4, 8 or 17 predictors (Table VI). It is important to note that the three- and four-predictor models contain continuous versions of the (rather strong) predictors age and Killip class. The three-predictor model is nested in the four-predictor model, and the eight-predictor model is nested in the 17-predictor model. The three- and four-predictor models are not nested in the eight- or 17-predictor models.

### *6.1. Total training part*

The performance of the models fitted in the total training part is considered as the gold standard. The gold standard four-predictor model performed slightly better than the three-predictor model; the model $\chi^2$ increased from 1597 to 1622. The eight- and 17-predictor models contained dichotomized versions of age ($>65$) and Killip class ($>2$), which caused a considerable loss of information. For example, age as a single continuous predictor had a model $\chi^2$ of 1032 in the training part, which decreased to 694 when dichotomized at 65 years. This explains why the gold standard performance of the eight- and 17-predictor models were rather similar to that of a three- or four-predictor model. The increase in model performance by adding nine, less powerful, predictors to the eight predictor was modest (model $\chi^2$ increased from 1604 to 1785).

### *6.2. Small subsamples*

The three-predictor model performed best in the small subsamples with 23 events on average (Table VI). The extension of this model with the predictor 'ST elevation' on average led to a slightly worse performance for all estimation methods. The performance of the eight-predictor models was relatively poor, especially when the standard ML estimates of the regression coefficients were used. This is explained by considerable overestimation of the coefficients, as indicated by the slope of the prognostic index (0.66 on average). This overestimation was also seen for the adapted estimates (0.81 on average), but much less with shrunk or penalized estimates (see also Figure 1). We further note that the variability of the model performance increased considerably when eight predictors were considered.

### *6.3. Large subsamples*

In the large subsamples, we observed a similar performance for the three- or four-predictor model (Table VI). The performance of the 17-predictor model was slightly worse than the eight-predictor

 

Table VI. Performance of three-, four-, eight- and 17-predictor models in the small and large subsamples (average $n = 336$ and $n = 892$, respectively), and in the total training part ($n = 20\,512$), as evaluated in the test part ($n = 20\,318$). Mean (standard deviation) shown for several estimation methods with a fixed selection of predictors.

| | Three predictors | Four predictors | Eight dichotomous predictors | 17 predictors |
|---|---|---|---|---|
| *Total training part* ('*gold standard*') (1423 *deaths*) | | | | |
| Area under ROC | 0.789 | 0.791 | 0.789 | 0.802 |
| Slope of prognostic index | 0.973 | 0.969 | 0.944 | 0.959 |
| Model $\chi^2$ | 1597 | 1622 | 1604 | 1785 |
| *61 small subsamples* (*23 deaths on average*) | | | | |
| Area under ROC | | | | |
| Standard/shrunk | 0.78 (0.017) | 0.77 (0.019) | 0.75 (0.029) | |
| Penalized | 0.78 (0.017) | 0.77 (0.018) | 0.76 (0.021) | |
| Adapted | 0.79 (0.003) | 0.79 (0.005) | 0.78 (0.027) | |
| Slope of prognositc index | | | | |
| Standard | 0.86 (0.19) | 0.82 (0.19) | 0.66 (0.18) | |
| Shrunk | 0.94 (0.22) | 0.94 (0.24) | 1.01 (0.29) | |
| Penalized | 0.97 (0.26) | 0.97 (0.26) | 0.93 (0.30) | |
| Adapted | 0.95 (0.11) | 0.91 (0.11) | 0.81 (0.19) | |
| Model $\chi^2$ | | | | |
| Standard | 1292 (383) | 1235 (391) | 673 (1211) | |
| Shrunk | 1339 (312) | 1309 (299) | 1045 (776) | |
| Penalized | 1324 (320) | 1309 (298) | 1112 (384) | |
| Adapted | 1552 (57) | 1526 (78) | 1261 (684) | |
| *23 large subsamples* (*62 deaths on average*) | | | | |
| Area under ROC | | | | |
| Standard/shrunk | 0.79 (0.003) | 0.79 (0.005) | 0.79 (0.009) | 0.78 (0.010) |
| Penalized | 0.79 (0.003) | 0.79 (0.005) | 0.78 (0.009) | 0.79 (0.009) |
| Adapted | 0.79 (0.001) | 0.79 (0.003) | 0.79 (0.002) | 0.79 (0.005) |
| Slope of prognostic index | | | | |
| Standard | 0.94 (0.15) | 0.91 (0.13) | 0.86 (0.13) | 0.76 (0.12) |
| Shrunk | 0.97 (0.15) | 0.97 (0.14) | 0.97 (0.16) | 0.95 (0.16) |
| Penalized | 0.98 (0.16) | 0.98 (0.15) | 0.96 (0.17) | 0.98 (0.19) |
| Adapted | 0.97 (0.09) | 0.96 (0.09) | 0.92 (0.08) | 0.86 (0.10) |
| Model $\chi^2$ | | | | |
| Standard | 1510 (102) | 1502 (123) | 1422 (120) | 1294 (277) |
| Shrunk | 1518 (89) | 1515 (105) | 1461 (95) | 1441 (163) |
| Penalized | 1514 (95) | 1515 (107) | 1455 (102) | 1497 (142) |
| Adapted | 1580 (73) | 1583 (41) | 1558 (35) | 1578 (135) |

* The three-predictor model is nested in the four-predictor model; these models are not nested in the eight- or 17-predictor models.

† The eight-predictor model contains dichotomous predictors only and is nested in the 17-predictor model (see text).

model when the standard or shrunk estimation was applied, and slightly better for penalized or adapted estimation. Hence, including more predictors did not clearly improve performance. The variability of the performance was large for the 17-predictor model. Further, we note that models constructed in the large subsamples were more stable than those constructed in the small

 

subsamples, which might be expected since the sample size was 2.7 times as large (62 compared to 23 events).

### *6.4. Regions*

The models listed in Table VI were also evaluated in the eight regions with 178 events on average. The four-predictor model performed similarly to the three-predictor model, but adding nine predictors improved the eight-predictor model (average model $\chi^2$ from 1537 to 1634, standard ML estimates). The best performance was noted for the 17-predictor model with adapted estimation of the coefficients, followed by penalized, shrunk or standard ML estimation (average model $\chi^2$ 1755, 1675, 1660 or 1634, respectively).

## 7. DISCUSSION

This study provides a number of insights to prognostic modelling in general, and logistic regression modelling in small data sets in particular. First, it emphasizes the importance of the use of external information in selection and estimation processes. This may not only improve the predictive performance of the model in future patients, but may also improve the clinical credibility of the model. Clinicians may be more inclined to apply a model which includes well-known predictors with a plausible sign [4, 31]. Second, the limited value of stepwise selection as a tool for prognostic modelling is confirmed [2, 10, 16]. The power for selection of important predictors may often be too low to make stepwise selection valuable; more information is lost than is gained. Third, shrinkage of regression coefficients may improve the performance of a prognostic model substantially. We found no major differences between application of a linear shrinkage factor, a penalized maximum likelihood procedure, or the Lasso. The Lasso is a promising technique, since shrinkage is defined such that some coefficients may be set to zero. The number of selected predictors was however quite large in our evaluations (for example, 16.3 of 17 predictors in samples with 62 events on average).

With modelling in a small data set, some general statistical principles deserve special attention. The data set is only a sample from an underlying population. Our aim is not to describe the sample as best as we can, but to learn about the population. We might expect that the more we study the data set, the more we learn about the population. This is only partly true with regression modelling. Data-driven decisions imply a better fit to the data under study, but much less so for the underlying population. This has been labelled the 'cost of data-analysis' [48, 49]. Examples of data-driven decisions include stepwise methods for selection of covariables (main effects, but also non-linear and interaction terms), univariable or graphical inspections, re-grouping of categorical variables based on the relation with the outcome, or the choice of an appropriate type of regression model based on the fit on the data. Data-driven model specification may make the apparent model performance severely overoptimistic [1–3, 15–17, 49].

In a small data set, we should not expect to be able to find a 'true' model, if such a model exists at all [14, 15, 50]. Rather a model should be specified based on external knowledge, which can be expected to describe the patterns in the data set sufficiently. If a large model is prespecified, stepwise selection may often not be able to help us in finding an adequate subset of this model.

 

A small data set has in this respect been defined as one with less than 10 events per variable (EPV) [31]. The 1:10 rule is somewhat arbitrary, and we may try to refine this criterion. First, it is evident that the total sample size is also important, in addition to the EPV. This was illustrated by the modelling of 17 predictors in data sets with 62 events, which was much less problematic than modelling eight predictors in data sets with 23 events, although the EPV value was only slightly larger (3.7 or 2.9). Second, we propose two additional critical EPV values: 20 and 50. When the 1:10 rule is violated, the number of parameters to be estimated may in fact be too large for the data under study. A small prespecified model should be fit with shrinkage of the regression coefficients. When the EPV is larger than 10 but smaller than 20, a prespecified model may adequately be fit, but shrinkage is advisable. When the EPV exceeds 20, shrinkage may not be necessary anymore for full models. Criteria for application of stepwise selection with $\alpha = 0.05$ are difficult to provide. We performed some additional analyses with 17-predictor models where nine covariables were made randomly associated with the outcome. These analyses indicated that stepwise selection did not improve predictive performance compared with shrunk full models unless EPV exceeded 50 (that is, only in the total training data set). Note that the number of candidate predictors should be considered in this reasoning, not the number of predictors included in the final model. It may hence often be impossible to study a comprehensive set of potential predictors, since this may easily amount to 50 to 100 predictors in prognostic problems. Further research of EPV criteria is indicated.

In our evaluations, we focused on the quality of the predictions resulting from a logistic regression model ('best predictions'). This focus should be distinguished from learning about the most important prognostic relationships ('best predictors'). Stepwise methods are attractive for the latter, although several drawbacks should be considered, such as instability of the selection, limited power and biased estimation of regression coefficients [5, 23]. Selection from a predefined set of predictors with external information (for example, plausibility of the sign) and a limited use of stepwise methods with a high $\alpha$ may be a reasonable compromise between predictive accuracy and insight in important predictive relationships.

Several limitations apply to our study. Foremost, the analyses with the GUSTO-I data represent essentially a case study. Although the structure of the data set may be representative of other clinical prediction problems, exceptions can probably be identified, for example, where covariables have stronger collinearity or stronger relative effects. Also, predictive factors for 30-day mortality after an acute MI have been widely studied, and models containing relevant predictors could be readily prespecified. When plausible prior predictive relationships are not known, prognostic modelling in a small data set will be harder. Next, heterogeneity of patient populations between centres or geographic locations was not taken into account in our analyses, while this may be an important consideration in clinical practice [51, 52]. Further, our consideration of selection and estimation methods was far from complete. We have only attempted to include techniques that are commonly used with logistic regression in the medical domain (backwards stepwise selection), or that may be used in the future (shrinkage techniques, quantitative or qualitative external information). Other modelling approaches may provide better predictions. We encourage comparative study of such methods and methods that look promising from our analysis (penalized ML, Lasso, adaptation, sign OK).

In conclusion, we have severe reservations regarding the routine use of stepwise selection for prognostic modelling in small data sets. Instead, full models should be considered, with shrinkage of the coefficients [1–3, 16]. External knowledge should be incorporated as much as possible in the modelling process.

 

ACKNOWLEDGEMENTS

We thank Kerry L. Lee and Amanda L. Stebbins, Duke Clinical Research Institute, Duke University Medical Center, Durham NC, for making the GUSTO-I data available for analysis, and Jørgen Hilden, Department of Biostatistics, University of Copenhagen, Denmark, and two anonymous reviewers for their comments. Part of this research was supported by a grant from the Netherlands Organization for Scientific Research (NWO, S96-156).

REFERENCES

1. Harrell F, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.
2. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 1986; **5**:421–433.
3. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models: *Statistics in Medicine* 1990; **9**:1303–1325.
4. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules: a review and suggested modifications of methodological standards. *Journal of the American Medical Association* 1997; **277**:488–494.
5. Miller AJ. *Subset Selection in Regression*. Chapman and Hall: London, 1990.
6. Akaike H. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Petrov BN, Csaki F (eds), Akademia Kiado: Budapest, 1973; 267–281.
7. Miller AJ. *Subset Selection in Regression*. Chapman & Hall: London, 1990.
8. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 1992; **42**:265–282.
9. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer* 1994; **69**:979–985.
10. Harrell FE, Lee K, Califf RM, Pryor DB, Rosati RA. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine* 1984: **3**:143–152.
11. Chen C-H, George SL. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine* 1985; **4**:39–46.
12. Altman DG, Andersen PK. Bootstrap investigation of the stability of the Cox regression model. *Statistics in Medicine* 1989; **8**:771–783.
13. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* 1992; **11**:2093–2109.
14. Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. *Biometrics* 1997; **53**:603–618.
15. Chatfield C. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*, *Series A* 1995; **158**:419–466.
16. Copas JB. Regression, prediction and shrinkage (with discussion) *Journal of the Royal Statistical Society*, *Series B* 1983; **45**:311–354.
17. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* 1983; **78**:316–331.
18. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: London, 1993.
19. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association Statistics in Medicine* 1992; **87**:942–951.
20. Verweij PJM, Van Houwelingen JC. Penalized likelihood in Cox regression. *Statistics in Medicine* 1994; **13**:2427–2436.
21. Hoerl AE, Kennard RW. Ridge regression: biased estimates for nonorthogonal problems. *Technometrics* 1970; **12**:55–67.
22. Bancroft TW, Ian CP. Inference based on conditional specification. *International Statistical Review* 1977; **45**:117–128.
23. Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*. 1999; **52**:935–942.
24. Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, *Series B* 267–288.
25. Tibshirani R. The Lasso method for variable selection in the Cox model. *Statistics in Medicine* 1997; **16**: 385–395.
26. Breiman L. Better subset regression using the nonnegative Garotte. *Technometrics* 1995; **37**:373–384.
27. Steyerberg EW, Kievit J, De Mol Van Otterloo JCA, Van Bockel JH, Eijkemans MJC, Habbema JDF. Perioperative mortality of elective abdominal aortic aneurysm surgery: a clinical prediction rule based on literature and individual patient data. *Archives of Internal Medicine* 1995; **155**:1998–2004.

 

28. Steyerberg EW, Eijkemans MJC, Houwelingen JC van, Lee KL, Habbema JDF. Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine* 2000; **19**:141–160.
29. Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews* 1987; **9**:1–30.
30. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Statistics in Medicine* 1993; **12**:717–736.
31. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996; **49**:1373–1379.
32. Concato JC, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Annals of Internal Medicine* 1993; **118**:201–210.
33. Harrell FE. Design: S-plus functions for biostatistical/epidemiological modelling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. Programs available at internet: http://lib.stat.cmu.edu/DOS/S/Harrell/,1997.
34. Lawless JF, Singhal K. Efficient screening of nonnormal regression models. *Biometrics* 1978; **34**:318–327.
35. Tibshirani R. Lasso. Programs available at internet: http://lib.stat.cmu.edu/S/,1995.
36. Steyerberg EW. Sample5.zip S + program and data set available at internet: http://www.eur.nl/fgg/mgz/software.html, 1998.
37. The GUSTO-I Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *New England Journal of Medicine* 1993; **306**:673–682.
38. Lee KL, Woodlief LH, Topol EJ, Weaver WD, Betriu A, Col J, Simoons M, Aylward P, Werf F van der, Califf RM, for the GUSTO-I investigators. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: results from an international trial of 41 021 patients. *Circulation* 1995; **91**: 1659–1668.
39. Steyerberg EW, Harrell FE, Goodman PH. Neural networks, logistic regression, and calibration: a rejoinder [letter]. *Medical Decision Making* 1998; **18**:445–446.
40. Mueller HS, Cohen LS, Braunwald E, Forman S, Feit F, Ross A, Schweiger M, Cabin H, Davison R, Miller D, Solomon R, Knatterud GL. Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction. *Circulation* 1992; **85**:1254–1264.
41. Dubois C, Pierard LA, Albert A, Smeets J-P, Demoulin J-C, Boland J, Kulbertus HE. Short-term risk stratification at admission based on simple clinical data in acute myocardial infarction. *American Journal of Cardiology* 1988; **61**:216–219.
42. Maggioni AP, Maseri A, Fresco C, Franzosi MG, Mauri F, Santoro E, Tognoni G, on behalf of the Investigators of the Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico (GISSI-2). Age-related increase in mortality among patients with first myocardial infarctions treated with thrombolysis *New England Journal of Medicine* 1993; **329**:1442–1448.
43. Maynard C, Weaver WD, Litwin PE, MJartin JS, Kudenchuk PJ, Dewhurst TA, Eisenberg MS, Hallstrom AP, Chambers J, for the MITI Project Investigators. Hospital mortality in acute myocardial infarction in the era of reperfusion therapy (the Myocardial Infarction Triage and Intervention project). *American Journal of Cardiology* 1993; **72**:877–882.
44. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Statistics in Medicine* 1991; **10**:1213–1226.
45. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Journal of the American Medical Association* 1982; **247**:2543–2546.
46. Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods of Information in Medicine* 1978; **17**:227–237.
47. Hinton GE. Connectionist learning procedures. *Artificial Intelligence* 1989; **40**:185–234.
48. Faraway JJ. On the cost of data analysis. *Journal of Computational and Graphical Statistics* 1992; **1**:213–229.
49. Ye J. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*. 1998; **93**:120–131.
50. Nester MR. An applied statistician's creed. *Applied Statistics* 1996; **45**:401–410.
51. Poses RM, Cebul RD, Collins M, Fager SS. The importance of disease prevalence in transporting clinical prediction rules. *Annals of Internal Medicine* 1986; **105**:586–591.
52. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine* 1995; **14**:1999–2008.

 