Statistica Neerlandica (2001) Vol. 55, nr. 1, pp. 76–88

Application of shrinkage techniques in logistic regression analysis: a case study

E. W. Steyerberg*, M. J. C. Eijkemans, J. D. F. Habbema

Center for Clinical Decision Sciences, Department of Public Health, Erasmus University, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands

Logistic regression analysis may well be used to develop a predictive model for a dichotomous medical outcome, such as short-term mortality. When the data set is small compared to the number of covariables studied, shrinkage techniques may improve predictions. We compared the performance of three variants of shrinkage techniques: 1) a linear shrinkage factor, which shrinks all coefficients with the same factor; 2) penalized maximum likelihood (or ridge regression), where a penalty factor is added to the likelihood function such that coefficients are shrunk individually according to the variance of each covariable; 3) the Lasso, which shrinks some coefficients to zero by setting a constraint on the sum of the absolute values of the coefficients of standardized covariables.

Logistic regression models were constructed to predict 30-day mortality after acute myocardial infarction. Small data sets were created from a large randomized controlled trial, half of which provided independent validation data. We found that all three shrinkage techniques improved the calibration of predictions compared to the standard maximum likelihood estimates. This study illustrates that shrinkage is a valuable tool to overcome some of the problems of overfitting in medical data.

Key Words and Phrases: regression analysis, logistic models, bias, variable selection, prediction.

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, and informing patients and their families (see e.g. HARRELL et al. 1996). The probability of a dichotomous medical outcome may well be estimated with a logistic regression model. An important problem is that medical data sets are often small compared to the number of covariables studied. Regression models constructed in such small data sets provide overconfident predictions in

* steyerberg@mgz.fgg.eur.nl

Part of this research was supported by a grant from the Netherlands Organization for Scientific Research (NWO, S96−156). Ewout Steyerberg is a fellow of the Royal Netherlands Academy of Arts and Sciences.

independent data: higher predictions will be found too high, and low predictions too low. COPAS (1983) and VAN HOUWELINGEN and LE CESSIE (1990) have proposed shrinkage techniques as a remedy against such extreme predictions.

In this study, we compare three shrinkage techniques for the estimation of logistic regression coefficients in small data sets: linear shrinkage, penalized maximum likelihood (or ridge regression), and the Lasso. For a general theoretical background of shrinkage, we refer to the paper of Van Houwelingen in this issue of *Statistica Neerlandica*. We here describe the predictive performance of logistic regression models which are constructed in small parts of a large data set of patients with an acute myocardial infarction to predict 30-day mortality. We only consider pre-specified models. In another publication, we described the effects of shrinkage in combination with various model specification techniques, especially stepwise selection (STEYERBERG et al., 2000).

We will first describe the three shrinkage techniques that we studied and their implementation (section 2). The patient data are described in section 3. Evaluations of predictive performance are presented in section 4. We discuss our findings in section 5.

2 Shrinkage techniques considered

We consider the usual logistic regression model $\text{logit}\{Y = 1|X\} = \beta_0 + \Sigma\beta_i \cdot X_i = PI$ where Y is a binary outcome variable (0 or 1), β_0 is an intercept, and β_i denotes the logistic regression coefficients for the design matrix X of covariables i. PI is the prognostic index, which is equivalent to the 'linear predictor' in the context of generalized linear models. Our aim is to estimate the $\text{logit}\{Y = 1|X\}$ accurately for new patients; interpretation of regression coefficients is secondary in our analyses.

Linear shrinkage factor

A relatively straightforward approach is to apply a linear shrinkage factor, s, for the regression coefficients as estimated with standard maximum likelihood (ML). According to COPAS (1983) and VAN HOUWELINGEN and LE CESSIE (1990), the shrunken coefficients (β_s) are then estimated as: $\beta_s = s \cdot \beta$. We estimated s with a bootstrap procedure:

1. Take a random bootstrap sample from the original sample, with the same size and patient records drawn with replacement (see e.g. EFRON, 1993).
2. Estimate the logistic regression coefficients in the bootstrap sample.
3. Calculate the PI for each patient in the original sample. The PI is the linear combination of the regression coefficients as estimated in the bootstrap sample with the values of the covariables in the original sample.
4. Estimate the slope of the PI with logistic regression, using the outcomes of the patients in the original sample and the PI as a single covariable.

The slope of the PI is by definition unity in step 3. In step 4 it will generally be smaller than 1, reflecting that regression coefficients are too extreme for predictive purposes. Steps 1 to 4 were repeated 300 times to obtain a stable estimate of s, which was calculated as the mean of the 300 slopes estimated in step 4 ($s =$ mean(slope)). Essentially, the expected miscalibration was estimated and used to correct the initially estimated regression coefficients. Linear shrinkage may be therefore referred to as 'shrinkage after fitting'.

Penalized ML

Ridge regression was proposed HOERL and KENNARD already in 1970 as a method to obtain less extreme regression coefficients, and has also been applied by LE CESSIE and VAN HOUWELINGEN (1992) with logistic regression. We use the more general term penalized maximum likelihood estimation (see e.g. HARRELL et al., 1996). For estimation of coefficients, a penalty factor λ is included in the maximum likelihood formula (see e.g. VERWEIJ and VAN HOUWELINGEN, 1994): $\log L - \frac{1}{2}\lambda\beta' P\beta$. Here L denotes the usual likelihood function, λ is the (positive) penalty factor, β' denotes the transpose of the vector of estimated regression coefficients β (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 0. 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 λ, we varied λ over a grid e.g. 0, 0.5, 1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64 and evaluated a modified Akaike Information Criterion (AIC): AIC $=$ [model $\chi^2 - 2 \cdot$ effective d.f.]. Here, model χ^2 is the likelihood ratio χ^2 of the model (i.e. compared to the null model with an intercept only), ignoring the penalty function. The effective degrees of freedom are calculated according to GRAY (1992): trace[$I(\beta)$cov(β)]. In the latter formula, $I(\beta)$ is the information matrix as computed without the penalty function, and cov(β) is the covariance matrix as computed by inverting the information matrix calculated with the penalty function. Note that if both the $I(\beta)$ and the cov(β) are estimated without penalty, $I(\beta)\,\mathrm{cov}(\beta)$ is the identity matrix and trace[$I(\beta)\,\mathrm{cov}(\beta)$] is equal to the number of estimated parameters in the model (excluding the intercept). With a positive penalty function, the elements of cov(β) become smaller and the effective degrees of freedom decrease. The λ with the highest AIC was used in the penalized estimation of the coefficients.

Note that although only a single shrinkage parameter is estimated (λ), a varying degree of effective shrinkage (s_e) is attained for individual coefficients. We may refer to penalized ML or ridge regression as 'shrinkage during fitting'.

Lasso

Another form of a penalized ML procedure is the Lasso method ('Least Absolute Shrinkage and Selection Operator') as proposed by TIBSHIRANI (1996). The Lasso combines shrinkage with selection of predictors, since some coefficients are shrunk to zero. It was developed in the same spirit as BREIMAN's Garotte (1995). We studied the Lasso since it can readily be applied to linear regression models but also to generalized linear models such as logistic regression or Cox proportional hazards regression. The Lasso estimates the regression coefficients β of standardized covariables while the intercept is kept fixed. The log-likelihood is minimized subject to $\Sigma|\beta| \leqslant t$, where the constraint t determines the shrinkage in the model. We varied $s = t/\Sigma|\beta^0|$ over a grid from 0.5 to 0.95, where β^0 indicates the standard ML regression coefficients and s may be interpreted as a standardized shrinkage factor. When $s = 1, \beta = \beta^0$ fulfills the constraint and no shrinkage is attained. We estimated β with the value of t that gave the lowest mean-squared error in a generalized cross-validation procedure. We may refer to the Lasso as 'shrinkage with selection'.

Estimation of required shrinkage

The described implementations of shrinkage techniques are available for S-plus software (MathSoft Inc., Seattle WA), with functions for linear shrinkage and penalized ML programmed by Harrell (http://lib.stat.cmu.edu/DOS/S/Harrell/) and lasso functions by Tibshirani (http://lib.stat.cmu.edu/S/). A drawback of these implementations is that the shrinkage parameters are estimated with rather different techniques. One would anticipate that any shrinkage parameter could be estimated with some form of cross-validation or bootstrapping. For the Lasso, fivefold cross-validation however gave poor results in TIBSHIRANI's simulation study (1996), but one might hypothesize that other variants of cross-validation (e.g. $20 \times$ fivefold) might work well. AIC and effective degrees of freedom can be applied for linear shrinkage and penalized ML. These two techniques are closely related; a linear shrinkage factor s is estimated in the penalized ML procedure when the matrix P is equal to the full matrix of second derivatives, with $s = 1/(1 + \lambda)$.

For our evaluation we used the most practical implementations of the shrinkage techniques: linear shrinkage by bootstrap resampling, penalized ML by AIC and effective degrees of freedom, and the Lasso by generalized cross-validation. The linear shrinkage factor might however have been estimated even simpler with Van Houwelingen and Le Cessie's heuristic formula: s_{heur} =[model χ^2 − (df − 1)]/model χ^2, where df indicates the degrees of freedom of the covariables in the model and model χ^2 is calculated on the log-likelihood scale. It is readily understood that s_{heur} approaches zero when larger numbers of predictors are considered (since the df increase), or when the sample size is smaller (since model χ^2 decreases).

3 Empirical evaluation

Patients

For evaluation of the shrinkage techniques we used the data of 40,830 patients with an acute myocardial infarction from the GUSTO-I study. LEE et al. (1995) decribed the data in detail, which were previously analysed by ENNIS et al. (1998) and

STEYERBERG et al. (1999, 2000). Mortality at 30-days was studied, which occurred in 2,851 patients (7.0%). Within the total data set, we distinguished 16 regions: 8 in the United States (US), 6 in Europe and 2 other (Canada and Australia/New Zealand).

The data set was split in a training and test part. These parts each consisted of eight regions with geographical balance and a similar overall mortality (7.0%). Within regions in the training part ($n = 20, 512$), 'large' and 'small' multicenter subsamples were created by grouping hospitals together on a geographical basis. We note that 'small' and 'large' are meant here in relation to the need for shrinkage, which is higher in smaller samples. 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. In this way, 61 small and 23 large subsamples of the training part were created, containing on average 336 and 892 patients of whom 23 and 62 died respectively. Logistic regression models were fitted in the subsamples and evaluated in the test part.

We note that the subsamples were not strictly random samples. Rather, it was aimed to replicate the real-life situation that a small multicenter data set would be available that contained patients from several nearby hospitals to construct a prognostic model which should be applicable to the total patient population.

Predictors considered

We considered two prognostic models. First, we focused on evaluations of an 8-predictor model, as defined by MUELLER et al. (1992) based for the TIMI-II study. This model included as predictors: shock, age >65 years, anterior infarct location, diabetes, hypotension, tachycardia, no relief of chest pain, and female gender. All 8 predictors were dichotomous, including age, where a continuous function would have been more appropriate. Second, we considered a 17-predictor model in the larger samples. This model consisted of the TIMI-II model plus nine other previously studied covariables (see e.g. LEE et al., 1995).

Table 1 shows the distribution of the 17 covariables and their uni- and multivariable logistic regression coefficients in the 8 and in the 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. 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 coefficients in the 8-predictor model were significant at the $p < 0.001$ level. The 17-predictor model contained covariables with relatively small coefficients, such as sex, hypertension, previous angina and family history.

Evaluation

For the evaluation of model performance we considered discrimination, calibration, and overall performance. Discrimination refers to the ability to distinguish high risk patients from low risk patients. In medical studies, discriminative ability is com-

Table 1. Distribution of predictors and logistic regression coefficients (standard error) in the 8-predictor and in the 17-predictor model. Results are shown for the total training part ($n = 20, 512$, 1,423 died) from the GUSTO-I data set

		Logistic regression coefficients		
Predictors	Prevalence¶	Univariable	8-pred. model	17-pred. model
Age> 65 year	41%	1.52	1.38 (0.06)	1.14 (0.07)
Female gender	24%	0.77	0.45 (0.06)	0.08 (0.09)
Diabetes	13%	0.58	0.27 (0.08)	0.29 (0.08)
Hypotension (BP < 100)	8%	1.28	1.24 (0.08)	1.25 (0.08)
Tachycardia (pulse > 80)	31%	0.75	0.67 (0.06)	0.65 (0.06)
Anterior infarct location	39%	0.93	0.76 (0.06)	0.43 (0.07)
Shock (Killip III/IV)	2%	2.51	1.74 (0.12)	1.69 (0.12)
No relief of chest pain	65%	0.55	0.52 (0.07)	0.53 (0.07)
Previous MI	16%	0.79		0.59 (0.07)
Height (·10 cm)*	17.1	−0.47		−0.16 (0.05)
Weight (·10 kg)*	7.9	−0.29		−0.11 (0.03)
Hypertension	37%	0.30		0.11 (0.06)
Smoking*#	1.87	0.48		0.17 (0.04)
Hypercholesterolaemia	35%	−0.27		−0.18 (0.07)
Previous angina	37%	0.40		0.14 (0.06)
Family history	41%	−0.37		−0.13 (0.06)
ST elevation in > 4 leads	37%	0.65		0.35 (0.07)

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

monly quantified by a concordance statistic (c) (see e.g. HARRELL et al., 1996). For binary outcome data 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 (see e.g. HILDEN et al., 1978). We used the slope of the prognostic index (PI), since this measure is readily interpretable in the context of shrinkage of regression coefficients (see e.g. MILLER et al., 1991). The PI 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. When regressing the observed outcome in the test part on the PI, the slope of the PI will usually be less than 1, indicating that low predictions are on average too low, and high predictions on average too high. We did not consider other aspects of goodness-of-fit in detail.

Finally, we aimed to quantify the overall model performance in one number. The model χ^2 is the difference in the -2log-likelihood with model predictions and the -2 log-likelihood of a model with an intercept only. The model χ^2 quantifies the agreement between predictions and observed outcomes and is similar to considering the Kullback–Leibler distance. It is a natural candidate to indicate overall performance, since it is measured on the same scale as the criterion which is being maximized in the model fitting procedure (-2 log-likelihood). The model χ^2 was calculated by fitting a model with an intercept and the prognostic index as an offset variable (slope fixed at unity, i.e. the prognostic index was taken literally) in the test data. A negative model χ^2 implied that a model performed worse than predicting the average risk for every patient.

4 Results in small data sets

Illustration: a small subsample

We first illustrate the use of the shrinkage techniques with a small subsample, which showed results that were typical for the other small subsamples. Table 2 shows the regression coefficients of the 8-predictor model 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 five Australian centers that participated in the GUSTO-I trial. The sample included 336 patients, of whom 20 died.

The 8-predictor model had large regression coefficients for age and shock when estimated with the standard ML procedure. The coefficients were shrunk with a factor 0.63 according to the bootstrapping procedure. Penalized estimates of the regression coefficients were obtained with a penalty factor of 8. The effective 'shrinkage' was 1.17 for shock, negative for anterior MI and relief of pain (which both had very small effects), and around 0.6 for the other covariables. This evaluation illustrates that the sign of the coefficients can be changed by the penalized ML procedure. The lasso parameter s was 0.78, which resulted in an increase of the estimated effect for shock, and a major shrinkage for anterior MI and relief of pain (shrunk to zero). As a reference, the final columns show the coefficients obtained in the total training part ('gold standard', $n = 20, 512$).

Predictions were calculated for the independent test part ($n = 20, 318$) according to each estimation method. Figure 1 shows the distribution of the prognostic index (logit of predicted probabilities). The standard ML estimates led to a broad range of predictions, which was drawn closer to the mean with linear shrinkage, penalized ML and the Lasso. We also note a slight shift in distribution to higher predicted logits for the Lasso. This is explained by inaccurate estimation of the intercept in this example.

Figure 2 shows the calibration of the predictions. We note that the standard ML estimates lead to a clear underestimation of the risk of death for low risk patients (e.g. probability $< 5\%$, logit < -3), and an overestimation of higher risks (e.g. predicted probability 40%, observed probability around 30%). All three shrinkage methods led to improved predictions, as indicated by curves closer to the identity line. This is also indicated by the slope of the PI, which was 0.68 for the standard ML estimates and around 1 with the shrinkage techniques (Table 2).

Shrinkage had only a minor advantage with respect to discriminative ability. Note that the c statistics for the models with standard or shrunk coefficients are by definition identical, since the ordering of the predicted probabilities does not change

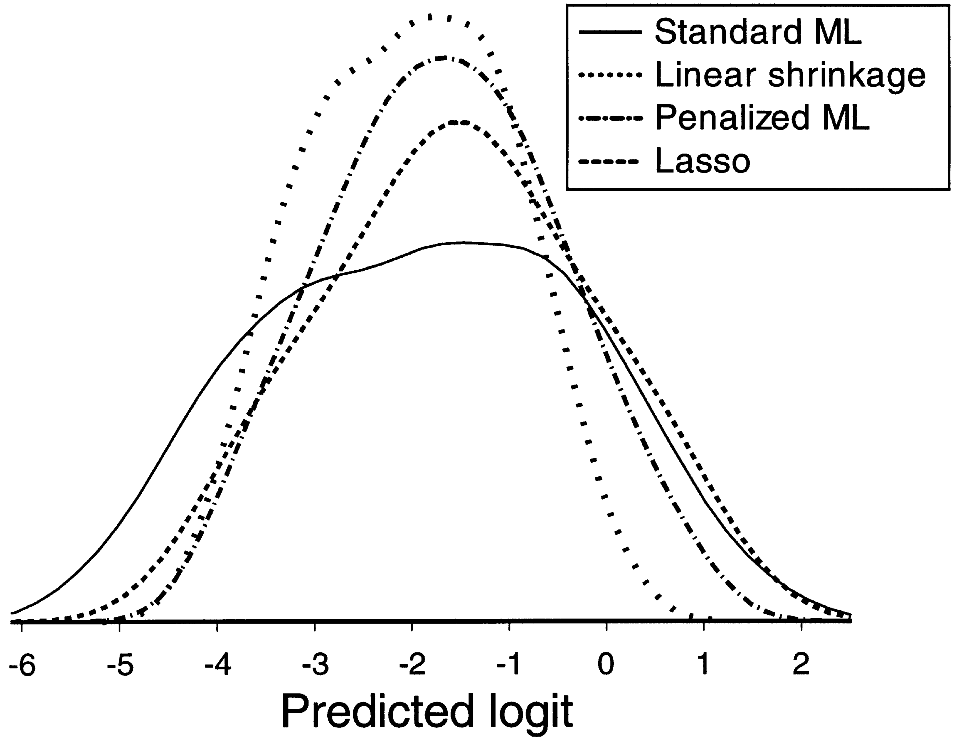

Fig. 1. Distribution of the predicted logit of mortality in the test part of the GUSTO-I trial ($n = 20, 318$). An 8-predictor model was fitted in a small subsample consisting of 336 patients of whom 20 died. Estimation of regression coefficients was with standard ML, a linear shrinkage factor, penalized ML, or the Lasso.

by applying a linear shrinkage factor. The shrinkage techniques led to some improvement in overall performance, as indicated by the model χ^2.

Average performance

We repeated the construction of a model in a subsample with testing in independent data for each of the 61 small and 23 larger subsamples. The low prevalence of some predictors led to zero cells and non-convergence of the coefficient estimates in the 8-predictor model for 12 of the 61 small subsamples. This problem might not have occurred had we considered convergence of the linear predictor rather than of the coefficients. The 12 non-converged subsamples were excluded from the evaluations. In Table 3, we show the average results of the performance of the estimated models as well as the performance obtained with a model fitted in the total training data set ($n = 20, 512$, 'gold standard').

The gold standard models had c statistics near 0.80, and good calibration (slopes around 0.95). Overall model performance increased slightly by adding nine predictors to the 8-predictor model (model χ^2 increased from 1604 to 1785).

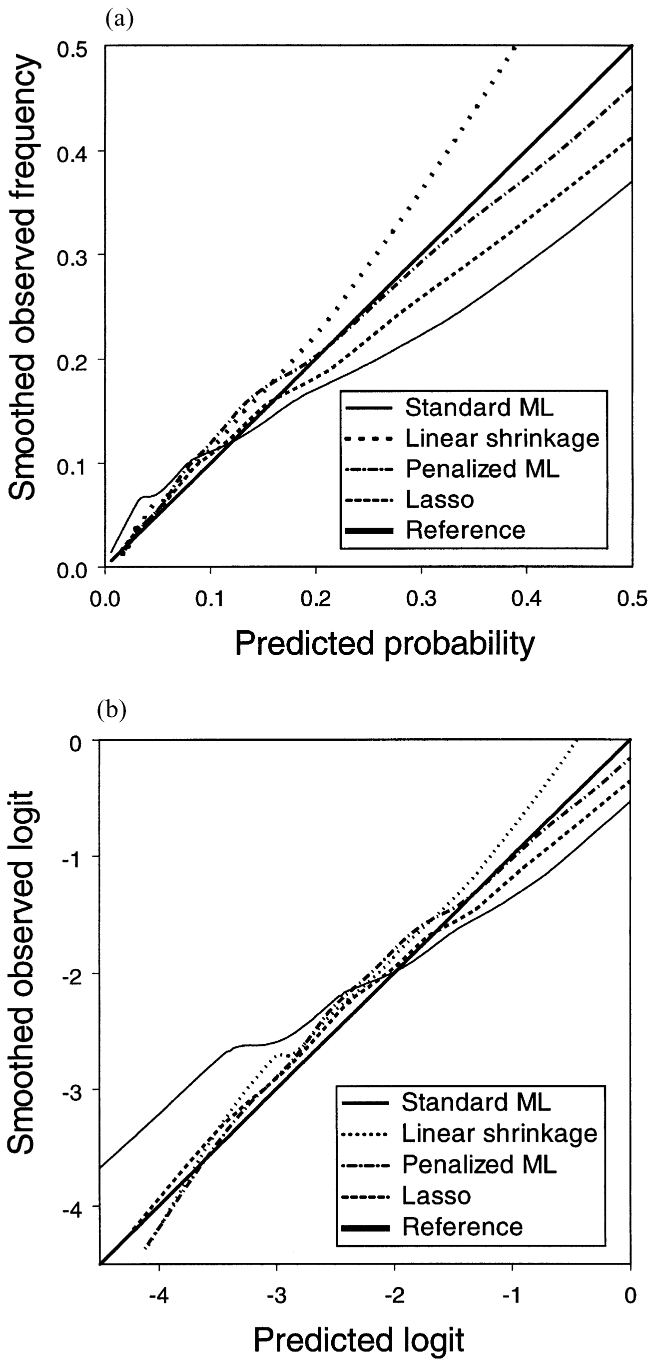

Fig. 2 Calibration plots of the 8-predictor model in the test part of the GUSTO-I trial ($n = 20, 318$). The model was fitted in a small subsample consisting of 336 patients of whom 20 died. Curves are non-parametric, constructed with the supersmoother algorithm in S+, and plotted on the probability scale (a) and the logit scale (b).

Table 2. Estimates of logistic regression coefficients (β) and effective shrinkage factor ($s_e = \beta/\beta_{\text{ML}}$) for the 8-predictor model in a small subsample (336 patients, 20 died) from the GUSTO-I trial, according to several estimation methods. Model performance was evaluated in the test sample (see text)

	Standard ML	Shrinkage methods						Gold standard
		Shrunk		Penalized		Lasso		
	β_{ML}	β	s_e	β	s_e	β	s_e	β_{Gold}
Predictors								
Shock	1.32	0.83	0.63	1.55	1.17	1.69	1.27	1.73
Age> 65 years	2.52	1.58	0.63	1.36	0.54	1.73	0.69	1.37
Anterior MI	−0.03	−0.02	0.63	0.08	−2.67	0	0	0.76
Diabetes	0.96	0.60	0.63	0.84	0.87	0.79	0.82	0.29
Hypotension	0.81	0.51	0.63	0.40	0.50	0.40	0.50	1.25
Tachycardia	o.91	0.57	0.63	0.59	0.64	0.65	0.71	0.66
No relief	0.02	0.01	0.63	−0.01	0.43	0	0	0.55
Female gender	0.19	0.12	0.63	0.30	1.55	0.15	0.81	0.44
Performance in test sample (n = *20, 318*)								
Area under ROC curve (c)	0.76	0.76		0.77		0.77		0.79
Slope of PI	0.68	1.09		1.01		0.90		0.94
Model χ^2	1082	1282		1322		1293		1604

Table 3. Average performance (standard deviation) of the 8 and 17-predictor model fitted 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 ($n = 20, 318$) from the GUSTO-I trial.

	8-predictor model			17-predictor model	
	Small	Large	Total	Large	Total
Area under ROC curve (c)					
Standard/Shrunk	0.754 (.029)	0.780 (.009)	0.789	0.777 (.010)	0.802
Penalized ML	0.760 (.021)	0.780 (.009)		0.784 (.009)	
Lasso	0.756 (.027)	0.784 (.009)		0.781 (.009)	
Slope of PI					
Standard	0.66 (0.18)	0.86 (0.13)	0.944	0.76 (0.12)	0.959
Shrunk	1.01 (0.29)	0.97 (0.16)		0.95 (0.16)	
Penalized ML	0.93 (0.30)	0.96 (0.17)		0.98 (0.19)	
Lasso	0.83 (0.23)	1.01 (0.16)		0.93 (0.15)	
Model χ^2					
Standard	673 (1211)	1422 (120)	1604	1294 (277)	1785
Shrunk	1045 (776)	1461 (95)		1441 (163)	
Penalized ML	1112 (384)	1455 (102)		1497 (142)	
Lasso	1079 (390)	1517 (84)		1492 (144)	

In the small subsamples, we found that using the standard ML estimates led to a poor overall performance. The area under the ROC curve was largely unaffected by applying penalized ML or the Lasso. The major improvement was seen with regard to calibration, where the slope of the prognostic index increased from 0.66 to values close to one for shrunk or penalized ML estimates, and 0.83 for the Lasso. The shrinkage techniques led to a better overall predictive performance, with model χ^2 over 1000 compared to 674 for standard ML estimates. Note however that the variability in performance was considerable.

In the larger subsamples, both the 8 and 17-predictor model were evaluated. The benefit of shrinkage was somewhat less in the larger samples. For the 8-predictor model, the average slope improved from 0.86 with standard ML estimates to values close to one with any of the shrinkage methods. However, when 17 predictors were considered, the average slope with standard ML estimates was 0.76, indicating a clearer need for shrinkage. Remarkably, the overall performance of the 17-predictor model was slightly worse than the 8-predictor model when coefficients were estimated with standard ML, linear shrinkage, or the Lasso, and marginally better with penalized ML. Hence, including more predictors did not clearly improve the performance. Further, we note that the performances of models constructed in the larger subsamples were more stable than in the small subsamples. This is consistent with the 2.7 times as larger sample size (on average 62 compared to 23 events).

5 Discussion

This study illustrates how shrinkage techniques can be applied with logistic regression analysis in small medical data sets. Shrinkage led to better calibration of predictions compared to models based on the standard ML estimates, especially when the data set was small compared to the number of covariables considered. We found no major differences in performance between application of a linear shrinkage factor, a penalized ML procedure similar to ridge regression, or the Lasso.

On the one hand, TIBSHIRANI's Lasso (1996) is an interesting technique, since shrinkage is defined such that some coefficients are set to zero. This leads to smaller predictive models, since covariables with a coefficient of zero can be dropped. Smaller models are more attractive for application by physicians in clinical practice. On the other hand, the number of predictors that was selected by the Lasso was quite large in our evaluations (e.g. 16.3 of 17 predictors in samples with 62 events on average). Also, calculation of the optimal Lasso parameter is computationally demanding, especially for larger data sets, and not attractive from a theoretical point of view. The theoretical foundation of penalized ML (or ridge regression) is stronger, with similarities in estimation with natural penalties and generalized additive models (see e.g. VAN HOUWELINGEN, 2001).

In medical prediction problems, an extensive model specification phase is quite common, where a model is sought that is eventually used in the final analysis. Model specification may include coding and categorization of covariables with 'optimal' cutpoints, determination of suitable transformations for continuous covariables, and selection of 'important' predictors (see e.g. CHATFIELD, 1995, for a critical discussion). Often, stepwise procedures are applied where covariables are omitted from (backward or stepdown selection) or entered in the model (forward selection) based on repeated significance testing. In a previous evaluation we found that omission of non-significant covariables decreased predictive performance, especially discrimination (STEYERBERG et al., 2000). Miscalibration could reasonably be corrected by all three shrinkage techniques considered in the present study. For linear shrinkage, the shrinkage factor was calculated in a bootstrapping procedure that included the stepwise selection process. Penalized coefficients were calculated with the same penalty as identified as optimal for the full model, although this approach lacks a theoretical foundation. Following CHATFIELD (1995), BUCKLAND et al. (1997), and YE (1998), we would argue that the model specification phase should not be ignored when interpreting regression coefficients in the final model or when predictions are based on the model.

Several limitations apply to our study. Foremost, the analyses with the GUSTO-I data represent a case study. Although the structure of the data set may be representative of other prediction problems in medicine, exceptions can probably be identified, e.g. where covariables have stronger collinearity or larger predictive effects. Further, we have only included implementations of a limited number of shrinkage techniques that may currently be used relatively easily with logistic regression. Bayesian and other recently proposed approaches were not included. We encourage further developments of shrinkage techniques, especially those that select variables by shrinking coefficients to zero or otherwise take selection of covariables properly into account.

Acknowledgements

We would like to thank Kerry Lee, Duke University Medical Center, Durham NC, and the GUSTO investigators for making the GUSTO-I data available for analysis; Frank Harrell, University of Virginia, Charlottesville VA, for valuable discussions; Hans van Houwelingen, Leiden University, Leiden, The Netherlands, and a reviewer for valuable comments.

References

BREIMAN, L. (1995), Better subset regression using the nonnegative Garotte, *Technometrics* **37**, 373–384.

BUCKLAND, S. T., K. P. BURNHAM and N. H. AUGUSTIN (1997), Model selection: an integral part of inference, *Biometrics* **53**, 603–618.

CHATFIELD, C. (1995), Model uncertainty, data mining and statistical inference, *Journal of the Royal Statistical Society, Series A* **158**, 419–466.

COPAS, J. B. (1983), Regression, prediction and shrinkage, *Journal of the Royal Statistical Society, Series B* **45**, 311–354.

EFRON, B. and R. TIBSHIRANI (1993), *An introduction to the bootstrap*, Monographs on statistics and applied probability, Vol. 57, Chapman & Hall, New York.

ENNIS, M., G. HINTON, D. NAYLOR, M. REVOW and R. TIBSHIRANI (1998). A comparison of statistical learning methods on the Gusto database, *Statistics in Medicine* **17**, 2501–2508.

GRAY, R. J. (1992), Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis, *Journal of the American Statistical Association* **87**, 942–951.

HARRELL, F. E., Jr., K. L. LEE and D. B. MARK (1996), Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine* **15**, 361–387.

HILDEN, J., J. D. HABBEMA and B. BJERREGAARD (1978), The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities, *Methods of Information in Medicine* **17**, 227–237.

HOERL, A. E. and R. W. KENNARD (1970), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics* **12**, 55–67.

LE CESSIE, S. and J. C. VAN HOUWELINGEN (1992), Ridge estimators in logistic regression, *Applied Statistics – Journal of the Royal Statistical Society, Series C* **41**, 191–201.

LEE, K. L., L. H. WOODLIEF, E. J. TOPOL, W. D. WEAVER, A. BETRIU, J. COL, M. SIMOONS, P. AYLWARD, F. VAN DE WERF and R. M. CALIFF (1995), Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients, *Circulation* **91**, 1659–1668.

MILLER, M. E., S. L. HUI and W. M. TIERNEY (1991), Validation techniques for logistic regression models, *Statistics in Medicine* **10**, 1213–1226.

MUELLER, H. S., L. S. COHEN, E. BRAUNWALD, S. FORMAN, F. FEIT, A. ROSS, M. SCHWEIGER, H. CABIN, R. DAVISON, D. MILLER, R. SOLOMON and G. L. KNATTERUD (1992), Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction, *Circulation* **85**, 1254–1264.

STEYERBERG, E. W., M. J. EIJKEMANS and J. D. HABBEMA (1999), Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis, *Journal of Clinical Epidemiology* **52**, 935–942.

STEYERBERG, E. W., M. J. EIJKEMANS, F. E. HARRELL, Jr. and J. D. HABBEMA (2000), Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets, *Statistics in Medicine* **19**, 1059–1079.

TIBSHIRANI, R. (1996), Regression and shrinkage via the Lasso, *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

VAN HOUWELINGEN, J. C. and S. LE CESSIE (1990), Predictive value of statistical models, *Statistics in Medicine* **9**, 1303–1325.

VAN HOUWELINGEN, J. C. (2001), Shrinkage and penalized likelihood as methods to improve predictive accuracy, *Statistica Neerlandica* **55**, 17–34.

VERWEIJ, P. J. and H. C. VAN HOUWELINGEN (1994), Penalized likelihood in Cox regression, *Statistics in Medicine* **13**, 2427–2436.

YE, J. (1998), On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association* **93**, 120–131.

Received: March 2000. Revised: July 2000.