Prognostic Modeling with Logistic Regression Analysis:

In Search of a Sensible Strategy in Small Data Sets

EWOUT W. STEYERBERG, PhD, MARINUS J. C. EIJKEMANS, MSc, FRANK E. HARRELL JR, PhD, J.DIK F. HABBEMA, PhD

Clinical decision making often requires estimates of the likelihood of a dichotomous outcome in individual patients. When empirical data are available, these estimates may well be obtained from a logistic regression model. Several strategies may be followed in the development of such a model. In this study, the authors compare alternative strategies in 23 small subsamples from a large data set of patients with an acute myocardial infarction, where they developed predictive models for 30-day mortality. Evaluations were performed in an independent part of the data set. Specifically, the authors studied the effect of coding of covariables and stepwise selection on discriminative ability of the resulting model, and the effect of statistical "shrinkage" techniques on calibration. As expected, dichotomization of continuous covariables implied a loss of information. Remarkably, stepwise selection resulted in less discriminating models compared to full models including all available covariables, even when more than half of these were randomly associated with the outcome. Using qualitative information on the sign of the effect of predictors slightly improved the predictive ability. Calibration improved when shrinkage was applied on the standard maximum likelihood estimates of the regression coefficients. In conclusion, a sensible strategy in small data sets is to apply shrinkage methods in full models that include well-coded predictors that are selected based on external information. ***Key words:*** regression analysis; logistic models; bias; variable selection; prediction. **(Med Decis Making 2001;21:45–56)**

Clinical decision making often requires estimates of the likelihood of an outcome in individual patients. For example, the concept of a treatment or test/treatment threshold requires that the probability of disease is estimated based on available patient characteristics and clinical findings.[1,2] Also, treatment choice may be guided by the risk of an adverse outcome, such as short-term mortality, where patients at greatest risk will gain the most.[3] When empirical data are available, individualized probability estimates for such outcomes may well be obtained from a logistic regression model, especially if such a model is developed in a large data set.[4]

Advantages of logistic regression include that few assumptions are made, for example, about the distribution of the outcome, and that interpretable results are provided, because the regression coefficients represent odds ratios.[5] The resulting predictive model is therefore no black box, in contrast to, for example, a neural network, which is inherently bound to the computer.

Neural networks have gained some interest in the medical decision-making literature, because better predictions might potentially be obtained by methods that pose fewer restrictions on the structure of the predictive model than logistic regression.[6–9] Statistically speaking, logistic regression is a generalized linear model. Without

Received 8 November 1999 from the Center for Clinical Decision Sciences, Department of Public Health, Erasmus University, Rotterdam, the Netherlands (EWS, MJCE, JDFH), and Division of Biostatistics and Epidemiology, Department of Health Evaluation Sciences, University of Virginia, Charlottesville, Virginia (FEH). Revision accepted for publication 9 August 2000. Ewout Steyerberg was supported by a grant from the Netherlands Organization for Scientific Research (NWO, S96-156) and a fellowship from the Royal Netherlands Academy of Arts and Sciences. Presented at the 1999 meeting of the Society for Medical Decision Making, October 4-6, Reno, Nevada.

Address correspondence and reprint requests to Dr. Steyerberg: Center for Clinical Decision Sciences, Ee2091, Department of Public Health, Erasmus University, Rotterdam, the Netherlands; telephone: 31-10-408 7053; fax: 31-10-408 9455; e-mail: steyerberg@mgz.fgg.eur.nl.

interactions, it may be considered a special case of a generalized nonlinear model, such as a neural network.[10,11]

When the predictive performance of a neural network model is compared to that of a logistic regression model, a sensible strategy should have been followed in the development of both.[12] Unfortunately, such strategies are not uniformly agreed upon. A sensible strategy for logistic regression modeling may borrow some of the strengths of neural network modeling, such as a concern for too extreme predictions and for overoptimistic estimates of predictive performance.

With neural networks, it is often attempted to prevent predictions from being too extreme by application of some sort of early stopping rule in the learning process.[9,13] A related principle for regression models is known as "shrinkage."[14,15] A shrinkage factor can empirically be estimated to reduce the coefficients in a regression model, such that better calibrated predictions are obtained.[5,15,16] Furthermore, the structure of a regression model may to a substantial extent be determined by the data. For example, statistically nonsignificant predictors may be excluded from the model with stepwise selection techniques and goodness-of-fit tests may be applied iteratively to make sure that the finally selected model fits the training data closely. This data-driven model specification will also contribute to too extreme predictions.[17,18] In contrast, neural networks are commonly specified a priori, for example, with 1 hidden layer; a priori specification is possible because this is not very restrictive.[9]

Overoptimistic estimates of predictive performance are a common problem of any predictive model.[16] The apparent performance will be higher than that in an independent sample of patients not considered in the modeling process. This problem holds especially in small data sets, that is, data sets with relatively few patients or outcomes in comparison with the number of candidate predictors.[19–21]

In this study, we explore the effects of commonly applied steps in the development of a logistic regression model. We focus on small data sets, in that we are often confronted with this situation when predictive models are developed in the context of a decision analysis. For model specification, we explore the effects of dichotomizing continuous variables, selection with stepwise methods from a full model including all candidate predictors, and selection of predictors based on the plausibility of their sign in the model. We also study the benefits of statistical shrinkage techniques. For our empirical evaluations, we use a large data set of patients with an acute myocardial infarction (MI) where we aim to predict 30-day mortality. We conclude with a proposal for a sensible modeling strategy in small data sets.

Patients and Methods

PATIENTS

For evaluation of the modeling strategies, we used the data of 40,830 patients with complete follow-up from the GUSTO-I clinical trial.[22] The large size of this data set makes it possible to create small subsamples, for which the results of modeling strategies can be evaluated on a large independent part of the data set. The GUSTO-I data set has been used before to develop a prognostic model[4,23] and to study the effects of modeling strategies.[13,18,24] In brief, this data set consists of patients with acute MI, who were randomized to 1 of 4 thrombolytic regimens. The differences between these regimens were small relative to the effect of predictive covariables, and they are ignored in the present analyses. Mortality at 30 days was the primary end point and occurred in 2851 patients (7.0%).

Within the total data set, we distinguished a training part and a test part (Figure 1). These parts each consisted of 8 regions with geographical balance, 4 in the United States, 3 in Europe, 1 other (Australia/New Zealand or Canada), and a similar overall mortality (7.0%). Within the training part (n = 20,512), 23 subsamples were created containing on average 892 patients of whom 62 died. The subsamples were created by grouping hospitals together on a geographical basis, such that at least 50 events occurred per subsample. The subsamples were not strictly random samples but aimed to reflect the real-life situation that a relatively small multicenter data set is available, which contains patients from several nearby hospitals to construct a prognostic model, which should be applicable to the total patient population. Logistic regression models were constructed in the subsamples from the training part and evaluated in the test part (n = 20,318). For illustration, we focused on 1 of the subsamples, containing 752 patients from 24 centers

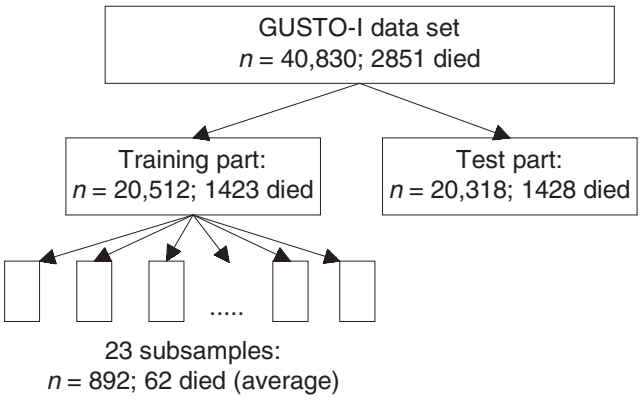

FIGURE 1. Graphical representation of the construction of a training part and a test part and 23 subsamples within the GUSTO-I data set.

in Australia ("example subsample"). This subsample was chosen because its results were representative for the pattern observed in the 23 subsamples. Furthermore, average results are presented over the 23 subsamples.

PREDICTORS CONSIDERED

We considered 3 previously developed prognostic models for acute MI as a basis for our evaluations. One study considered a 3-predictor model that contained the variables age and Killip class (a measure for left ventricular function) and the dichotomous variable anterior infarct location.[25] Age and Killip class were modeled as interval variables. In the GISSI-2 data set, the continuous variable "number of leads with ST elevation" was selected in addition to these 3 variables.[26] We further considered an 8-predictor model, as defined in the TIMI-II study.[27] This model included 8 dichotomous covariables: shock, age > 65 years, anterior infarct location, diabetes, hypotension, tachycardia, no relief of chest pain, and female gender. The effect of dichotomization of age and Killip class as > 65 and shock (approximately Killip class III/IV) on the predictive performance was evaluated.

In addition to these 3 models, a 17-predictor model was considered, consisting of the TIMI-II model plus 9 other covariables considered in previous analyses.[25–29] These 9 additional covariables included demographics (weight and height), history characteristics (previous MI, previous angina pectoris, hypertension, hypercholesterolemia, smoking, family history of MI), and ST elevation in more than 4 leads on the electrocardiogram (ECG).

Finally, we aimed to study the effect of including noise covariables in a prognostic model. We therefore created a new data set, where the values of each of the latter 9 covariables in the 17-predictor model were randomly permuted. This procedure generated 9 covariables without a true relationship with the outcome, but with a distribution identical to the 9 predictors in the original data set.

The number of events per variable (EPV) has been suggested as a criterion for the size of a data set, with small data sets having EPV < 10.[19–21] The subsamples that we considered contained on average 62 deaths, leading to EPV values of 21, 16, 7.8, and 3.6 for models with 3, 4, 8, and 17 predictors, respectively.

SELECTION METHODS

We applied stepwise selection with backward elimination of predictors from a full 17-predictor model. We initially used the standard selection criterion of $P < 0.05$ for inclusion of a covariable. We explored the effect of using a more liberal criterion by including only those covariables with $P < 0.5$. We further studied the effect of including interaction terms in the model (i.e., testing the assumption of additivity of predictors on the log odds scale) and nonlinear terms for the important continuous predictor age (i.e., testing the assumption of linearity on the log odds scale). Nonlinear terms included simple transformations (age^2, $age^{0.5}$, age^{-1}, log(age)), and a restricted cubic spline function with 3 knots (second-degree functions). This function provides slightly more flexibility in form than the previous terms and may hence be preferred when assessing nonlinearity.[30]

Furthermore, we applied a selection strategy that assumed qualitative knowledge on the sign of the predictive effect, which we labeled "Sign OK selection." The qualitative knowledge should be provided by uni- and multivariable analyses published before, or by clinical experts. In our study, we used the univariable signs from the training part of the GUSTO-I data set (n = 20,512) as our source of information. These signs were identical to the multivariable signs in both the training part and the test part, and therefore represent knowledge of the true signs of the predictive effects. In reality, knowledge of the signs may be imperfect. We therefore applied a variant of Sign OK selection, labeled "Conservative sign OK"

selection. Here, the covariables hypercholesterolemia, family history, and smoking were assumed to have a positive association with 30-day mortality, whereas in fact they had not. This variant is only one example of imperfect knowledge of the signs, whereas many other possibilities exist. Sign OK selection has some similarity with "Bayes-empirical-Bayes" estimation[31] and may prevent what epidemiologists have labeled the Type III error, that is, the inclusion of covariables with an incorrect sign.[19]

EVALUATION OF PERFORMANCE

The evaluation of model performance focused on discrimination and calibration.[5,32] Discrimination refers to the ability to distinguish high-risk patients from low-risk patients and is commonly quantified by a measure of concordance, the *c* index.[33] 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.[15,32,34] We used 1 simple measure to quantify calibration, that is, the slope of the prognostic index, which was originally proposed by Cox.[35] The slope of the prognostic index (or linear predictor) is the regression coefficient β in a logistic model with the prognostic index as the only covariate: observed mortality = α + β prognostic index.[32] The observed mortality is coded binary (0/1), and the prognostic index is calculated as the linear combination of the regression coefficients as estimated in a subsample with the values of the covariables for each patient in the test data. The slope of the prognostic index (referred to as the "calibration slope") should ideally be 1, when predicted risks agree fully with observed frequencies. Models providing overoptimistic predictions will show a slope that is less than 1, indicating that low predictions are too low and high predictions are too high.

VALIDATION PROCEDURE

Internal validity was assessed by bootstrapping.[36] Bootstrapping is a resampling method that allows one to make inferences about the population that the sample originated from by drawing with replacement from the original sample. A bootstrap sample may include a patient 0, 1, 2, . . . , *n* times, where *n* is the size of the original sample. Modeling strategies were validated as follows.

1) Take a random bootstrap sample from the original sample, identical in size and drawn with replacement.
2) In the bootstrap sample, select the covariables according to the selection procedure, estimate the logistic regression coefficients, and calculate performance measures (*c* index and calibration slope). This step provides insight in the variability of the selection process and provides estimates of the apparent performance.
3) Evaluate the model as estimated in the bootstrap sample in the original sample. The difference between the performance estimated in the bootstrap sample and in the original sample is an almost unbiased estimate of the overoptimism that may be expected for the model estimated on the original sample.[16,36]

Steps 1 to 3 were repeated 200 times to obtain stable estimates of the mean and standard error of the mean. Overoptimism was calculated as the average difference in performance (*c* index and calibration slope) between steps 2 and 3. For example, when the average *c* was 0.80 in step 2 and 0.78 in step 3, the overoptimism was 0.02. Furthermore, calibration slope was, by definition, unity in step 2. The average calibration slope in step 3 will generally be smaller than unity, reflecting overestimation of the regression coefficients for predictive purposes.[5,14,15] This reduced slope can be used as a linear shrinkage factor to improve predictions in future patients. Shrunken regression coefficients are then calculated by multiplication with the shrinkage factor.[15] For example, when the average slope in step 3 was 0.82, the regression coefficients would be multiplied by 0.82 for the final predictive model. The intercept in the final model should subsequently be adjusted such that predictions agree on average with the observed frequency of the outcome ("calibration in the large").

SOFTWARE

All calculations were performed with S-plus software (version 3.3, Data Analysis Products Division of MathSoft, Inc., Seattle, WA), where we

Table 1 • Discriminative Ability of Logistic Regression Models Developed in a Subsample (n = 785; 52 deaths) and Evaluated in a Test Part (n = 20,318; 1428 deaths) of the GUSTO-I Data with 3, 4, 8, or 17 Predictors, Including Age and Killip Class as Dichotomized or Continuous Covariables

		Number of Predictors			
	Coding of Age and Killip Class	3	4	8	17
Development data	Dichotomized	0.738	0.750	0.790	0.807
	Continuous	0.799	0.802	0.828	0.832
Test data	Dichotomized	0.745	0.739	0.780	0.784
	Continuous	0.787	0.784	0.805	0.805

Note: Numbers represent the c index (area under the receiver operating characteristic curve).

used Harrell's Design library.[37] An example of an S-plus program including the various modeling approaches is freely available.[38]

Results

CODING AND NUMBER OF PREDICTORS

The previously published prognostic models considered in our study differ with respect to their coding, especially of two strong predictors (i.e., age and Killip class, continuous or dichotomized) and with respect to the number of covariables (i.e., 3 to 17). In Table 1, we show c indices (areas under the ROC curve) as estimated in our example subsample ("development c," n = 752) and as observed in the test part of the GUSTO-I data set ("test c," n = 20,512) for the models considered, with continuous or dichotomized versions of age and Killip class.

As expected, dichotomization implied a loss of discriminative ability: c decreased by around 0.05 in the models with 3 or 4 predictors (Table 1). Models with 3 or 4 predictors and continuous versions of age and Killip class performed similar to 8- or 17-predictor models and dichotomized versions of these covariables (development c around 0.80, test c around 0.78).

Furthermore, we note that a fairly reasonable model could be obtained by considering only 3 predictors with continuous coding (test c = 0.787). The discriminative ability in the development data increased when 3, 4, 8, or 17 predictors were considered, respectively, but it reached a plateau at around 0.80 in the test data. The difference between the apparent and test performance is a reflection of the "overoptimism" that is seen when a model is evaluated on the same data set that was used to develop the model. The overoptimism was small for the 3-predictor model but increased to over 0.02 for the 17-predictor model.

The effects of dichotomization and inclusion of more covariables were confirmed in the analyses of the other subsamples. The test c in the test data was 0.786 for the 3- or 4-predictor models with

Table 2 • Logistic Regression Coefficients for 17 Predictors of 30-Day Mortality after Acute Myocardial Infarction (MI)

		Training Part	Subsample				
Label	Predictors	17-Predictor Model	17-Predictor Model	$P < 0.05$	$P < 0.50$	Sign OK	Conservative Sign OK
A65	Age > 65 years	1.14	1.08	1.15	1.09	1.03	1.18
SEX	Female sex	0.08	–0.75	—	—	—	—
DIA	Diabetes	0.29	0.42	—	0.44	0.42	0.46
HYP	Hypotension (BP < 100)	1.25	1.06	—	1.05	0.93	0.93
HRT	Tachycardia (pulse > 80)	0.65	1.05	0.98	1.03	1.01	0.94
ANT	Anterior infarct location	0.43	0.59	—	0.55	0.56	0.55
SHO	Shock (Killip III/IV)	1.69	1.14	1.32	1.15	1.14	1.10
TTR	No relief of chest pain	0.53	0.59	—	0.57	0.62	0.60
PMI	Previous MI	0.59	0.68	0.72	0.61	0.65	0.65
HEI	Height (•10 cm)[a]	–0.16	–0.30	—	–0.30	–0.07	–0.09
WEI	Weight (•10 kg)[a]	–0.11	–0.28	–0.26	–0.28	–0.26	–0.24
HTN	Hypertension	0.11	0.12	—	—	0.08	0.06
SMK	Smoking[a,b]	–0.17	–0.24	—	–0.24	–0.21	—
LIP	Hypercholesterolemia	–0.18	–0.01	—	—	–0.09	—
PAN	Previous angina	0.14	–0.15	—	—	—	—
FAM	Family history	–0.13	–0.38	—	–0.37	–0.40	—
ST4	ST elevation in > 4 leads	0.35	–0.08	—	—	—	—

Note: Results are shown for the training part (n = 20,512; 1423 deaths) and a subsample of the GUSTO-I data (n = 785; 52 deaths), where several selection strategies are considered.

a. Continuous predictor, modeled as linear term in logistic regression analysis.

b. Smoking was coded as 3 for current smokers, 2 for ex-smokers, 1 for never smokers.

Table 3 • Discriminative Ability of Logistic Regression Models Shown in Table 2 in the Test Part of the GUSTO-I Data Set (n = 20,318)

	Training Part	Subsample				
	17 Predictors	17-Predictor Model	$P < 0.05$	$P < 0.50$	Sign OK	Conservative Sign OK
Development data	0.805	0.807	0.777	0.807	0.805	0.802
Test data	0.802	0.784	0.762	0.784	0.795	0.790

Note: Numbers represent the c index (area under the receiver operating characteristic curve).

continuous versions of age and Killip class, and 0.780 and 0.777 for the 8- and 17-predictor models with dichotomous versions, respectively (averages over 23 subsamples).

INFLUENCE OF SELECTION

We first studied the regression coefficients in the full 17-predictor model and in models that were constructed according to the selection methods considered in this study (Table 2). The predictors age, hypotension, and shock had the strongest prognostic effects in the total training part (coefficients > 1).

In the subsample, stepwise selection with $P < 0.05$ as the selection criterion led to the exclusion of 12 predictors, leaving age, tachycardia, shock, previous MI, and weight in the reduced model. A substantially higher P value for selection ($P < 0.50$) led to the exclusion of 5 of the 17 predictors. Selection on the plausibility of the sign of the coefficient led to the exclusion of female sex, previous angina, and ST elevation as predictors. To explore the effect of selecting on the wrong sign, we considered a model where hypercholesterolemia, family history, and smoking were assumed to have a positive association with 30-day mortality ("Conservative Sign OK," Table 2).

In Table 3, we show the discriminative ability of the various models in the development and the test data. The model from the total training part obtained a c of 0.802 in the test data; this performance is considered a gold standard. The c for the full model as developed in the example subsample decreased by 0.023 (from 0.807 to 0.784). With $P < 0.05$ selection, the decrease was smaller but the test performance was worse (decrease 0.015; from 0.777 to 0.762). Hence, despite a smaller decrease in apparent performance, the test performance of the $P < 0.05$ stepwise selected model was worse than that of the full model. The best test performance was obtained with the "Sign OK" selected model (c = 0.795). Excluding some predictors, which had in fact the correct sign, led to a small decrease (c = 0.790). This performance was still better than that of the full 17-predictor model (c = 0.784).

These results from the subsample of 785 patients were confirmed in the other subsamples. The best performance in the test data was obtained with Sign OK selection (average c = 0.786). Intermediate results were obtained with $P < 0.50$ selection (average c = 0.771) or full models (average c = 0.777), and poorest with $P < 0.05$ selection (c = 0.758).

EXTENSION WITH NONLINEAR/INTERACTION TERMS

After selection of main effect terms in a prognostic model, it may be assessed whether nonlinear or interaction terms should be included. In general, tests for interaction require large samples to provide adequate power. As an example, we here focus on statistical interactions with age. In the total training part, a strong interaction was found between age and hypotension (χ^2 = 11, P = 0.001) and between age and Killip class (χ^2 = 8, P < 0.005). However, the P values for these interactions were 0.92 and 0.91, respectively, in the subsample. Another interaction with age (age and tachycardia) was significant in the subsample (P = 0.016), but not in the total training part (χ^2 = 1.3, P = 0.25). Because we usually do not have strong beliefs about interactions a priori, an overall test for interaction may provide insight into the need for interaction terms in the model. This test (16 df) had a P value of 0.52 in the subsample, indicating that no important second-order interactions with age were present. In the total training part, statistical significance was achieved ($P < 0.001$), but adding the 2 most important interaction terms did not improve the performance in the test part (c = 0.802 with or without interactions).

Furthermore, we investigated nonlinearity in age, starting with the 3-predictor model, which included age, Killip class, and anterior infarct. Simple nonlinear terms (age^2, $age^{0.5}$, age^{-1}, log(age)) and a

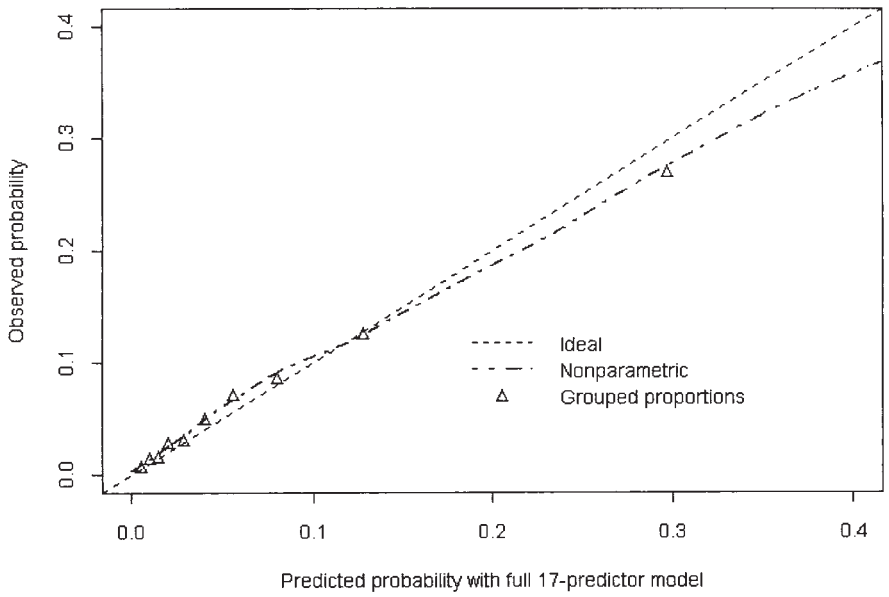

FIGURE 2. Calibration of predictions according to the 17-predictor model in the test set. The "ideal" line indicates perfect calibration of predictions. The observed calibration is shown with a smoothed curve ("nonparametric") and in deciles of predicted risk ("grouped proportions").

restricted cubic spline of age were far from statistically significant ($P > 0.8$), in agreement with findings in the training part, where the nonlinearity in age was relatively unimportant (main effect of age: $\chi^2 = 1374$, nonlinearity $\chi^2 = 16$). The extension of models with nonlinear or interaction terms was not explored further in the other subsamples, where we limited ourselves to models with main effects.

INFLUENCE OF NOISE COVARIABLES

In the previous analyses, all covariables were predictive of 30-day mortality. This may explain the relatively poor performance of the stepwise selected models with $P < 0.05$; stepwise selection could only lead to a loss of information. We might hope that this selection method would perform better if some of the candidate predictors were unrelated to mortality. We therefore also evaluated models that were constructed and tested in a data set where 9 of the 17 predictors were made randomly associated with the outcome. The *c* index of such a 17-predictor model was 0.785 in the test part when constructed in the total training part. This implies that the 9 noise covariables only moderately deteriorated model performance ($c = 0.802$ with true predictors).

In the example subsample with 785 patients, the full 17-predictor model had an apparent and test *c* of 0.801 and 0.753, respectively. Stepwise selection with $P < 0.05$ led to exclusion of all 9 noise covariables, but also to exclusion of 5 true predictors. This stepwise selected model included only age, shock, and tachycardia, with an apparent *c* of 0.749, which was close to the test *c* of 0.746. This performance was worse than that obtained with a full 17-predictor model (test $c = 0.753$), but the overoptimism was much smaller (0.003 versus 0.048). Sign OK selection could be applied in several ways. If we assume knowledge of the true signs, the 9 noise covariables would be excluded (sign = 0). This implies that we would consider the 8-predictor model, where the test *c* was 0.780 (Table 1). If we assume that the signs were similar to those used before (e.g., based on the opinion of a physician), we would select all 8 true predictors and 4 of the 9 noise covariables. This would result in a test *c* of 0.769, slightly exceeding that of the full model ($c = 0.753$).

These findings were confirmed by evaluations of the other subsamples. The test *c* was slightly poorer for the $P < 0.05$ stepwise models compared with the full 17-predictor models (average $c = 0.755$ and 0.759, respectively), with a somewhat better test *c* for Sign OK selection assuming similar sign as before (average $c = 0.771$). Note that the $P < 0.05$ stepwise models excluded 8.5 (expected 95% • 9 = 8.55) of the 9 noise covariables on average but at the same time 3.2 of the 8 true predictors. In contrast, Sign OK selection excluded on average 0.3 of the true predictors and 4.9 (expected 50% • 9 = 4.5) of the noise covariables. The exclusion of true predictors apparently was more detrimental for

Table 4 • Bootstrap Investigation of the Stability of the Backward Stepwise Elimination Process with $P < 0.05$, Which Resulted in the Inclusion of 5 Covariables as Shown in Table 2

	SHO	A65	ANT	DIA	HYP	HRT	TTR	SEX	PMI	HEI	WEI	HTN	SMK	LIP	PAN	FAM	ST4
Factors in final model																	
	*	*				*			*		*						
Factors retained in backwards elimination																	
	*								*	*							
	*	*	*			*	*										
	*	*	*			*											
	*	*							*							*	
	*	*		*		*					*						
	*	*	*	*	*		*	*	*		*	*					
	*	*		*	*	*	*	*		*			*				
		*	*			*				*				*	*		
	*	*	*		*	*	*						*				
	*	*				*					*					*	
	*	*			*		*		*								
		*				*	*						*				
		*				*		*			*						
		*				*		*	*		*	*				*	
	*	*				*							*				
	*	*				*			*		*					*	
	*	*	*		*	*					*		*				
	*		*						*		*						
		*	*			*					*					*	*
		*		*	*	*										*	*

Note: The results of 20 bootstrap samples are shown. SHO = Shock (Killip III/IV); A65 = Age; ANT = Anterior infarct location; DIA = Diabetes; HYP = Hypotension (BP < 100); HRT = Tachycardia (pulse > 80); TTR = No relief of chest pain; SEX = Female sex; PMI = Previous MI; HEI = Height (•10 cm); WEI = Weight (•10 kg); HTN = Hypertension; SMK = Smoking; LIP = Hypercholesterolemia; PAN = Previous angina; FAM = Family history; ST4 = ST elevation in > 4 leads.

discriminative ability than the inclusion of noise covariables.

CALIBRATION OF PREDICTIONS

In Figure 2, we illustrate the calibration of predictions in the test part for the 17-predictor model based on the example subsample (second column, Table 2). We note that high predictions are too high and low predictions are somewhat too low. The calibration slope was 0.85 (SE = 0.02) for this model. By definition, the slope was 1.0 when evaluated in the subsample itself, indicating an overoptimism in slope of around 15%.

Consideration of all 23 subsamples indicated that the miscalibration was very similar for the full 17-predictor models, stepwise selected models, or Sign OK models. The average slopes of the prognostic indices in the test part were 0.76 for the full 17-predictor models; 0.79 or 0.77 for $P < 0.05$ or $P < 0.50$ stepwise selected models, respectively; and 0.80 for Sign OK selected models. With 9 noise covariables included, the average slopes were of similar magnitude (0.74, 0.81, 0.74, and 0.81, respectively).

VALIDATION

Bootstrap resampling was used to study the internal validity of the modeling strategies. The variability of the selection process with backward stepwise selection (criterion: $P < 0.05$) is illustrated in Table 4, where we show the selected variables in the first 20 of 200 bootstrap samples, which were drawn with replacement from the subsample. The final model included 5 of the 17 candidate predictors (Table 2). Although these 5 predictors were the most common among those selected in the bootstrap samples, the exact 5-predictor model was selected in only 1 of the 200 bootstrap samples. It was selected in none of the other 22 subsamples.

Tables 1 and 3 showed that the apparent discriminative ability of the models was overoptimistic of the performance in independent data. For the example subsample, the overoptimism

could partly be predicted by bootstrapping. Bootstrapping estimated a decrease in *c* from 0.807 to 0.752 for the full 17-predictor model, and a decrease from 0.777 to 0.723 for the $P < 0.05$ stepwise model. This expected performance was a bit too pessimistic but correctly indicated that the $P < 0.05$ stepwise selected model would perform worst. For the 23 subsamples, bootstrapping predicted the average decrease in performance correctly.

Bootstrapping could also be used to predict the miscalibration. The predicted slope of the prognostic index was 0.79 for the subsample, which comes near the test slope of 0.85 in Figure 2. We note that the estimate of 0.79 had considerable uncertainty associated with it, as reflected in an SE of 0.10. The predicted slope was used as a uniform shrinkage factor by multiplication with the regression coefficients in the final model for each of the 23 subsamples. The miscalibration then nearly disappeared: the average slopes were close to 1 in the test data for full models, $P < 0.05$ or $P < 0.5$ stepwise, or Sign OK selected models (average slopes 0.95, 0.94, 0.94, 1.01, respectively).

Discussion

This study illustrates the effects of different modeling strategies on the performance of logistic regression models in new patients. Below, we discuss our empirical findings in the GUSTO-I data set, followed by a proposal for a sensible modeling strategy in small data sets.

EMPIRICAL FINDINGS

The first step in the modeling process is the selection and definition of covariables for the regression model. Our findings indicate that dichotomization of continuous covariables may cause a substantial loss of information. Furthermore, stepwise selection with a standard criterion ($P < 0.05$) did not improve the predictive performance compared to a model including all candidate covariables, even when over half of the covariables actually represented noise. So the contention that stepwise selection may decrease overfitting and hence improve predictive performance was incorrect. In our evaluations, selection on the plausibility of the sign of the predictor improved model performance. Extension with nonlinear terms or interaction was not beneficial in this prediction problem.

Our findings with respect to stepwise selection are in concordance with other studies.[20,31,39] It is remarkable that all investigators who have studied the statistical properties of this technique have concluded that it has severe deficiencies.[40–45] Despite this criticism, stepwise methods are widely used nowadays. Their popularity may stem from the property of providing small, readily interpretable models. Indeed, models with a limited number of predictors are more easy to apply in clinical practice, especially if presented in an attractive way.[21] Disadvantages of stepwise methods include instability of the selection of predictors (Table 4) and overestimation of regression coefficients.[17,18] The power for selection of predictors will often be too low, leading to a loss of predictive information from the model. This same problem occurs with the testing of statistical significance of more complex terms (interactions, nonlinear terms).

Selection with a higher *P* value (e.g., 0.50) may be a sensible alternative to obtain a smaller model. The high *P* value increases the power for selection of true predictors, excluding only those covariables with small effects, and limits the bias in selected coefficients.[18] In our evaluation, the gain in information by inclusion of true predictors did outweigh the loss caused by erroneous inclusion of noise covariables.

We further illustrated the need for shrinkage of the regression coefficients to improve calibration. A simple shrinkage method is the use of a linear shrinkage factor, based on the estimated miscalibration in a bootstrapping procedure.[14,15] More advanced shrinkage methods apply a penalty factor in the maximum likelihood formula.[46,47] Recently developed methods formulate the penalty in such a way that some coefficients are shrunk to zero.[48–50] We previously found that these methods achieve selection in a statistically correct way.[24] Also, Bayesian methods have been proposed,[17,45] which may especially be useful to express the uncertainty in predictions correctly, by averaging over alternative model specifications.[51] Also, shrinkage of regression coefficients results if a null effect is assumed as prior.

Limitations of our empirical evaluation include that this experience in GUSTO-I is essentially a case

Table 5 • A Suggested Strategy for Prognostic Modeling in Small Data Sets

Modeling Phase	Guidelines
Selection of predictors	
Classification	Published or clinically plausible classifications; grouping of related variables blinded to outcome
Inclusion	Quantitative and qualitative information from published studies; clinical knowledge; limited testing of main effects
Extension with nonlinear or interaction terms	Clinical knowledge; consideration of overall significance; limited testing of individual terms
Estimation of regression coefficients	Shrinkage of main effects (linear shrinkage or penalization); stronger shrinkage of nonlinear or interaction terms if inclusion based on statistical testing
Evaluation of model performance	Bootstrapping including any selection processes and other data-driven decisions
Presentation of model	Nomogram/table/prognostic score chart/meta-model

study, which limits the generalizability of our findings. A specific characteristic is that predictors for mortality after acute MI have widely been studied, which will not be the case for many other medical prognostic problems. On the other hand, our evaluations of the 17-predictor model may reflect the common situation where some covariables have a strong relationship with the outcome (e.g., age, shock, hypotension) and other covariables have weaker effects. Furthermore, mortality in the GUSTO-I database could adequately be described with a logistic model with main effects only.[13] Inclusion of interaction or nonlinear terms could hence not be expected to improve model performance. In other situations, such terms may be more relevant.

A SENSIBLE STRATEGY?

Some general advice related to prognostic modeling has previously been published.[5] In Table 5, we propose a modeling strategy to obtain accurate (i.e., well-calibrated and discriminating) predictions from relatively small data sets, for example, the situation that the number of events is less than 10 times the degrees of freedom of the covariables (EPV < 10). These guidelines are based on the average findings in the GUSTO-I data as described in this article, but also on other modeling studies[20,31,39] and on practical considerations.

In small data sets, we should be cautious in basing the structure of our model solely on the data under study.[17,45] The first aspect is the classification of variables, for example, the grouping of categorical variables, or the categorization of continuous variables. Regrouping of categorical variables should be blinded to the outcome. Previously published or clinically practical classifications are preferable over classifications that fit the data best. Also, related variables may be grouped as 1 covariable, again blinded to the outcome.[5] The continuous character of a predictor should in general be maintained to maximize predictive ability.[30]

Inclusion of covariables as main effects in the prognostic model should predominantly be based on external information. A systematic review of published studies may help to recognize important prognostic factors. The sign of the effect of a predictor in the literature may also be used for selection, provided that it is reasonable to make assumptions about the signs in the multivariable analysis. In addition, experienced clinicians may indicate which patient characteristics are important and have plausible signs. The set of candidate predictors should be kept small, since a well-chosen small model may perform better than a large model. Testing with stepwise selection is discouraged, especially with the traditional P value of 0.05. The use of a substantially higher P value (e.g., 0.50) may be considered to limit the loss of information.

For continuous variables, a linear effect is usually assumed, and inclusion of main effects implies additivity of the effects of the covariables on the log odds scale. Clinical knowledge may guide the use of nonlinear functions; we prefer flexible ones such as restricted cubic splines.[5,30] Similarly, biologically plausible interactions may be included. Such prespecified nonlinear and interaction terms might be retained in the final model irrespective of their statistical significance. An alternative approach might be to perform an overall score test for all possible 2nd-order interaction terms or all possible 2nd-order nonlinear terms. If significant, the most important individual interaction/nonlinear terms might be included, but the estimation of regression coefficients should subsequently take

this data-driven selection into account by applying a stronger shrinkage or penalization of the estimated coefficients. In general, we propose a conservative attitude toward the inclusion of nonlinear or interaction terms, in that average patterns in the data will be represented appropriately with main effects.[52]

Calibration of predictions from a logistic model will improve by some form of shrinkage of the regression coefficients. A linear shrinkage factor can be estimated with the output from many software packages ((model $\chi^2 - k$)/model χ^2, where k is the degrees of freedom of the covariables in the model[14,15]), whereas other calculations require specialized software.[5,37]

Furthermore, selection and estimation may be considered as only the first 2 phases in the development of a prognostic model, followed by evaluation of model performance and presentation. Evaluation should generally include resampling techniques (e.g., bootstrapping or cross-validation) to assess internal validity.[5,16,36] Any data-driven decisions should be accounted for in the evaluation.[17,45,51] This may not be easy in practice, because techniques like bootstrapping can only incorporate automatic modeling decisions such as stepwise selection.[40] If we wish that the model be readily applicable in clinical practice, we should not present regression coefficients with their confidence intervals only.[21] For example, a nomogram,[53–55] a table with predicted outcome probabilities,[56] or a prognostic score chart[57] may be constructed. Furthermore, we may consider the description of a complex model with a simpler "meta-model," which includes a smaller number of predictors.[58]

Conclusions

When a small data set is analyzed with logistic regression to provide individualized estimates of the probability of a disease or an adverse outcome in a decision problem, it is crucial that a sensible strategy is followed in the development of the model. The validity of predictions from a logistic model will often be poor when the predictors in the model are stepwise selected with the default P value of 0.05 and the standard maximum likelihood estimates are applied. Instead, full models should be considered, with shrinkage of the coefficients. External knowledge should be incorporated as much as possible in the modeling process.

The authors wish to thank Kerry L. Lee, Amanda L. Stebbins, Duke Clinical Research Institute, Duke University Medical Center, Durham, NC, and the GUSTO investigators for making the GUSTO-I data available for analysis. The reviewers provided helpful comments on an earlier version of this manuscript.

References

1. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med. 1975;293:229–34.
2. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980;302:1109–17.
3. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. BMJ. 1995;311:1356–9.
4. Califf RM, Woodlief LH, Harrell FE Jr, et al. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J. 1997; 133:630–9.
5. Harrell F, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
6. Tu JV, Weinstein MC, McNeil BJ, Naylor CD. Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? The Steering Committee of the Cardiac Care Network of Ontario. Med Decis Making. 1998;18:229–35.
7. Lapuerta P, L'Italien GJ, Paul S, et al. Neural network assessment of perioperative cardiac risk in vascular surgery patients. Med Decis Making. 1998;18:70–5.
8. Orr RK. Use of a probabilistic neural network to estimate the risk of mortality after cardiac surgery. Med Decis Making. 1997;17:178–85.
9. Penny W, Frost D. Neural networks in clinical medicine. Med Decis Making. 1996;16:386–98.
10. Hastie T, Tibshirani R. Generalized additive models for medical research. Stat Methods Med Res. 1995;4:187–96.
11. Goodman PH. NevProp Software, version 4. Reno, NV: University of Nevada, 1998. http://www.scs.unr.edu/nevprop/
12. Steyerberg EW, Harrell FE Jr, Goodman PH. Neural networks, logistic regression, and calibration [Letter]. Med Decis Making. 1998;18:349–50.
13. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the gusto database. Stat Med. 1998;17:2501–8.
14. Copas JB. Regression, prediction and shrinkage (with discussion). J R Stat Soc B. 1983;45:311–54.
15. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med. 1990;9:1303–25.
16. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc. 1983; 78:316–31.
17. Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc A. 1995;158:419–66.
18. Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52: 935–42.
19. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49: 1373–9.

20. Harrell FE, Lee K, Califf RM, Pryor DB, Rosati RA. Regression modeling strategies for improved prognostic prediction. Stat Med. 1984;3:143–52.
21. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules: a review and suggested modifications of methodological standards. JAMA. 1997;277:488–94.
22. The GUSTO-I Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. N Engl J Med. 1993;306:673–82.
23. Lee KL, Woodlief LH, Topol EJ, et al., 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: 1659–68.
24. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF. Selection and shrinkage in logistic regression analysis. Stat Med. 2000; in press.
25. Dubois C, Pierard LA, Albert A, et al. Short-term risk stratification at admission based on simple clinical data in acute myocardial infarction. Am J Cardiol. 1988;61:216–19.
26. Maggioni AP, Maseri A, Fresco C, et al., 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. N Engl J Med. 1993;329:1442–8.
27. Mueller HS, Cohen LS, Braunwald E, et al. Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction. Circulation. 1992;85:1254–64.
28. Maynard C, Weaver WD, Litwin PE, et al., for the MITI Project Investigators. Hospital mortality in acute myocardial infarction in the era of reperfusion therapy (the Myocardial Infarction Triage and Intervention project). Am J Cardiol. 1993;72:877–82.
29. Selker HP, Griffith JL, Beshansky JR, et al. Patient-specific predictions of outcomes in myocardial infarction for real-time emergency use: a thrombolytic predictive instrument. Ann Intern Med. 1997;127:538–56.
30. Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst. 1988;80:1198–202.
31. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. Stat Med. 1993;12:717–36.
32. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med. 1991;10:1213–26.
33. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247: 2543–6.
34. Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis: II. Trustworthiness of the exact values of the diagnostic probabilities. Methods Inf Med. 1978;17:227–37.
35. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45:562–5.
36. Efron B, Tibshirani RJ. An introduction to the bootstrap. London: Chapman & Hall, 1993.
37. 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, 1997. Programs available at: http://www.lib.stat.cmu.edu/S and hesweb.med.virginia.edu/ biostat/s/Design.html.
38. Steyerberg EW. Sample4.zip. S+ program and data set. Available at: http://www.eur.nl/fgg/mgz/software.html, 1999.
39. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med. 1986;5:421–33.
40. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset algorithms: frequency of obtaining authentic and noise variables. Br J Math Stat Psychol. 1992;42:265–82.
41. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer. 1994;69:979–85.
42. Chen C-H, George SL. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. Stat Med. 1985;4:39–46.
43. Altman DG, Andersen PK. Bootstrap investigation of the stability of the Cox regression model. Stat Med. 1989; 8:771–83.
44. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. Stat Med. 1992;11:2093–109.
45. Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. Biometrics. 1997;53:603–18.
46. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc. 1992;87:942–51.
47. Verweij PJM, Van Houwelingen JC. Penalized likelihood in Cox regression. Stat Med. 1994;13:2427–36.
48. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc B. 1996;58:267–88.
49. Tibshirani R. The Lasso method for variable selection in the Cox model. Stat Med. 1997;16:385–95.
50. Breiman L. Better subset regression using the nonnegative Garotte. Technometrics. 1995;37:373–84.
51. Draper D. Assessment and propagation of model uncertainty. J R Stat Soc B. 1995;57:45–97.
52. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. Epidemiology. 1997; 8:429–34.
53. Lubsen J, Pool J, van der Does E. A practical device for the application of a diagnostic or prognostic function. Methods Inf Med. 1978;17:127–9.
54. Pryor DB, Harrell FE, Lee KL, Califf RM, Rosati RA. Estimating the likelihood of significant coronary artery disease. Am J Med. 1983;75:771–80.
55. Spanos A, Harrell FE, Durack DT. Differential diagnosis of acute meningitis: an analysis of the predictive value of initial observations. JAMA. 1989;262:2700–7.
56. Steyerberg EW, Keizer HJ, Messemer JE, et al. Residual pulmonary masses following chemotherapy for metastatic nonseminomatous germ cell tumor: prediction of histology. Cancer. 1997;79:345–55.
57. Steyerberg EW, Keizer HJ, Fosså SD, et al. Prediction of residual retroperitoneal mass histology following chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from 6 study groups. J Clin Oncol. 1995;13:1177–87.
58. Harrell FE Jr, Margolis PA, Gove S, et al. Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants. WHO/ARI Young Infant Multicentre Study Group. Stat Med. 1998;17:909–44.