Urine Cation (nmol/day) Supplementary Table 4 Toyohara T. et al.

Appl. Statist. (1999)
48, *Part* 3, *pp.* 313–329

The use of resampling methods to simplify regression models in medical statistics

Willi Sauerbrei

University of Freiburg, Germany

[Received January 1996. Final revision November 1998]

Summary. The number of variables in a regression model is often too large and a more parsimonious model may be preferred. Selection strategies (e.g. all-subset selection with various penalties for model complexity, or stepwise procedures) are widely used, but there are few analytical results about their properties. The problems of replication stability, model complexity, selection bias and an overoptimistic estimate of the predictive value of a model are discussed together with several proposals based on resampling methods. The methods are applied to data from a case–control study on atopic dermatitis and a clinical trial to compare two chemotherapy regimes by using a logistic regression and a Cox model. A recent proposal to use shrinkage factors to reduce the bias of parameter estimates caused by model building is extended to parameterwise shrinkage factors and is discussed as a further possibility to illustrate problems of models which are too complex. The results from the resampling approaches favour greater simplicity of the final regression model.

Keywords: Backward elimination; Bootstrap; Cross-validation; Model complexity; Prediction; Selection bias; Selection level

1. Introduction

In fitting regression models data analysts are often faced with many predictor variables which may have an influence on an outcome variable. In the following $\mathbf{X} = (X_1, \ldots, X_k)$ denotes the vector of predictor variables under consideration and $g(\mathbf{X}) = \beta_1 X_1 + \ldots + \beta_k X_k$ a linear function. A linear regression model is given by

$$E(Y|\mathbf{X}) = \beta_0 + g(\mathbf{X}) \qquad \text{with } \operatorname{var}(Y|\mathbf{X}) = \sigma^2.$$

For a binary outcome the logistic regression model,

$$\log\{P(Y = 1|\mathbf{X})/P(Y = 0|\mathbf{X})\} = \beta_0 + g(\mathbf{X}),$$

is used. For possibly censored survival times, Cox's proportional hazards model (Cox, 1972) is usually chosen. The influence of predictors is modelled through the hazard rate

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp\{g(\mathbf{X})\},$$

with $\lambda_0(t)$ denoting an unspecified base-line hazard rate. If the number of variables is large, a parsimonious model involving fewer than k variables may be preferable. An aim of the analysis is the selection of the subset of 'important' predictors, i.e. the variables whose regression coefficients β_j differ from 0 in the function $g(\mathbf{X})$. For this task, sequential strategies, such as forward, stepwise or backward procedures, or all-subset selection with different

Address for correspondence: Willi Sauerbrei, Institute of Medical Biometry and Medical Informatics, University of Freiburg, Stefan-Meier-Strasse 26, 79104 Freiburg, Germany.
E-mail: wfs@imbi.uni-freiburg.de

 0035–9254/99/48313

optimization criteria, such as Mallows's C_p, the Akaike information criterion AIC or the Bayesian information criterion BIC are available in many statistical packages and are often used in practical applications (Miller, 1990). Despite the importance of variable selection strategies and enormous attention in the statistical literature, many open issues remain concerning their properties. Comparisons of selection strategies are usually made on specific data sets or by using simulation (Roecker, 1991; Breiman, 1992; Sauerbrei, 1992).

Usually, methods for variable selection and related issues have been developed and investigated for a linear regression model. However, the methods, or at least their basic ideas, are commonly transferred to generalized linear models and to models for survival data, although there are sometimes additional problems, such as definitions of residuals or equivalents of R^2 for censored data.

In what follows we assume applications with several potential predictors (10–25) and at most moderate sample sizes (100–1000) and that the first steps in model building, such as decisions about categorization and coding of variables or the specification of the functional relationship between a predictor and the outcome variable, are considered carefully; the importance of these decisions is often underestimated and only partly described in applications.

The data analyst then has the problem of selecting those variables that should enter the 'final' regression model. A good model should fit the data adequately, not be too complex and serve as an accurate predictor for new observations. This leads to the central question: *how many variables* should enter the model (Miller (1990), pages 169–205)? Including variables without influence (unimportant variables) in the final model leads to the problem of overfitting and of unnecessarily increasing the variance of parameter estimates for correlated predictors. Also the variance of the prediction will be increased. Eliminating influential variables from the final model results in underfitting of the data with a worsening of the fit (increased residuals) and usually in biased parameter estimates for other variables (the 'omission bias'; Miller (1990), p. 173). The effect of this bias depends on the (multi-) collinearity between the omitted variables and the selected variables. Adding a variable to a given model may increase the goodness of fit, but it increases the complexity of the model and may improve or worsen its predictive ability. Depending on the aim of a particular study the severity of the potential biases induced by model building may be considered in different ways. When interest is only in a predictive model, the main concern may be its overoptimism. A 'complex' model may be tolerated and the contribution of each individual variable is hardly considered; the corresponding standard errors may be large. In medical applications the effect of a single variable, and whether it should be included in a final model, is often of central interest. Commonly several variables are considered in a multivariable regression model and multicollinearity between them may cause severe problems, which may be reduced by the elimination of some variables. Ideally this should result in (nearly) independence of the factor of main interest from the remaining factors in the model; however, this is often not achievable. This issue must be taken into consideration by interpreting the final fitted model. The differences between these two situations with main interest on the predictive model or interest on the effect of a single variable were stressed by Copas (1983), who stated that the loss functions are different and that

> 'a good predictor may include variables which are not significant, exclude others which are, and may involve coefficients which are systematically biased'.

With increasing computational power, resampling methods such as cross-validation and bootstrapping (Efron and Tibshirani, 1993) are becoming more popular for investigating problems caused by data-driven variable selection. Although the theoretical basis is often not

well developed, these computer-intensive approaches do offer one possibility for such investigations where results based on asymptotic theory are scarce.

Two data sets are used in a discussion of methods of investigating problems introduced by variable selection: one from a case–control study with 23 predictors, analysed with a logistic regression model; the other from a clinical trial with 15 predictors and the survival time as the outcome. Using backward elimination (BE) with different selection levels, cross-validation and bootstrap resampling, emphasis is placed on the choice between simple and complex final models. Furthermore a cross-validation approach to estimate (global) shrinkage factors is extended to parameterwise shrinkage factors (PWSFs). The difference of this new approach is illustrated in the two studies. The investigations with resampling methods help to demonstrate severe problems of selected final models and are used to argue in favour of greater simplicity. In Section 2 we briefly introduce variable selection methods and in Section 3 the two examples. Section 4 discusses problems caused by variable selection and some approaches based on resampling methods to handle them, including the extension to PWSFs. In Section 5 we give the results of the investigations using the resampling methods proposed. Section 6 discusses the importance of the selection level and of resampling strategies in the complex process of model building.

2. Variable selection methods

In the linear regression model several assessments of all-subset strategies for variable selection have been proposed (C_p, AIC and BIC), which use the error sum of squares as the measure of goodness of fit and a penalty term for model complexity. A detailed review of these, and many other model selection criteria which are often only minor modifications of them, is given in Teräsvirta and Mellin (1986).

Sequential procedures, such as forward selection, stepwise selection and BE, are established as the other important variable selection methods. A common feature of these methods is that starting from the null model (forward selection or stepwise selection) or from the full model (BE) a new variable is added or deleted in each step until a prespecified selection level, in software packages sometimes called 'p-value to enter' or 'p-value to remove', is reached. Analytical results about the true selection level are not available, but Sauerbrei (1992, 1993) concluded from simulation studies that, at least in the linear regression model, the true selection levels of the stepwise approaches are only slightly higher than the prespecified levels. In what follows BE(p) denotes variable selection with BE and a predefined selection level p for elimination and reinclusion. In text-books the importance of the selection level is usually not discussed in detail. Because of asymptotic and simulation results on the selection level for the all-subsets criterion and the simulation results for the stepwise approaches, Sauerbrei (1992) argued that BE(0.157) may be used as a substitute for all-subsets procedures with C_p or AIC. Models selected with BE(0.157) and AIC usually have at most minor differences (Sauerbrei, 1992; Blettner and Sauerbrei, 1993). BE is available in many statistical software packages for logistic and Cox regression models. Therefore BE(0.157) is used in this paper as a proxy for AIC; also BE(0.05) and BE(0.01) are used to investigate model complexity.

3. Two case-studies

3.1. Atopy study

In an unmatched case–control study 345 patients with chronically relapsing flexual eczema were diagnosed with 'atopic dermatitis'. As a control group 618 employees without a history

of eczema from the army and from the same university hospital were investigated. In addition to age and sex, 19 binary atopic features, such as personal history of allergic asthma, white dermographism, xerosis or food intolerance, were assessed independently by two experienced physicians. On the basis of their assessment and agreement, the variables were classified as 'subjective' or 'objective' (Table 1). Two laboratory measures (immunoglobulin E and a screening test for inhalant allergy) were also investigated and classified as 'normal' or 'elevated'. For further details see, for example, Diepgen *et al.* (1996). The parameter estimates with corresponding standard errors in the full logistic regression model indicate some strong effects and some variables which seem to have no influence on the case–control status. Using the logistic regression model and BE as the selection strategy, diagnostic scores with high discriminative ability should be developed. These scores are necessary to standardize the diagnosis of atopic dermatitis for practical use, and for clinical or epidemiological studies. For wider use a further score should be developed which does not require the laboratory measurements.

3.2. Glioma study

A randomized trial to compare two chemotherapy regimes included 447 patients with malignant glioma. At the time of the analysis 293 patients had died and the median survival

Table 1. Atopy study: parameter estimates $\hat{\beta}$ and standard errors SE for the full model M_1 and for the variables selected by BE (selection level 0.01)†

Variable	*Estimates for all patients*				*Estimates for subset I: ID even*		*Estimates for subset II: ID odd*	
	M_1		M_2		M_3		M_4	
	$\hat{\beta}$	*SE*	$\hat{\beta}$	*SE*	$\hat{\beta}$	*SE*	$\hat{\beta}$	*SE*
Intercept	−7.12	1.00	−7.73	0.69	−9.51	1.42	−8.02	1.01
sex	−0.24	0.45						
age	−0.05	0.03						
X_3	1.79	0.45	1.79	0.41	2.60	0.68	2.56	0.59
X_4	1.76	0.44	1.78	0.39	3.17	0.72		
X_5	2.95	0.75	2.71	0.65	3.85	1.14	4.06	0.91
X_6	0.99	0.46	1.15	0.39				
X_7	1.12	0.49	1.26	0.46				
X_8	2.58	0.50	2.51	0.45	3.41	0.82	2.47	0.60
X_9	2.99	0.50	2.74	0.44	2.70	0.72	4.06	0.73
X_{10}	2.41	0.53	2.21	0.48	3.00	0.86	3.36	0.69
X_{11}	−0.48	0.77						
X_{12}	0.60	0.46					1.98	0.57
X_{13}	0.30	0.44						
X_{14}	0.07	0.58						
X_{15}	−0.89	0.57						
X_{16}	0.81	0.52			2.22	0.70		
X_{17}	1.89	0.51	1.72	0.41			2.12	0.55
X_{18}	3.86	0.73	3.64	0.65	4.12	1.17	4.60	0.86
X_{19}	1.50	0.43	1.58	0.40	2.01	0.67		
X_{20}	2.02	0.48	1.95	0.43	3.97	0.86		
X_{21}	−0.23	0.49						
X_{22}	1.04	0.55			2.42	0.83		
X_{23}	0.12	0.62						

†The selected model M_2 is based on all patients and models M_3 and M_4 for two subsets determined by the patients' number ID. X_3–X_{15} are objective variables, X_{16} and X_{17} are laboratory measurements and X_{18}–X_{23} are subjective variables.

time from the date of randomization was about 11 months. Besides therapy, 12 variables (age, three ordinal and eight binary variables) which might influence the survival time were considered. The three variables measured on an ordinal scale (the Karnofsky index, the type of surgical resection and the grade of malignancy) were each represented by two dummy variables, resulting in a total of 15 predictors denoted by $X_1, \ldots, X_{15}$. For these predictors complete data were available for 413 patients (274 events) used here in a complete case analysis. Because of the investigation of shrinkage factors described in the next section, the predictors are standardized to mean 0 and standard deviation 1. Standardization has no influence on the results of the variable selection procedures used here. The regression parameter estimates with corresponding standard errors in the full model indicate that some strong predictors are present (the upper part of Table 2). A detailed description of the study and a comparison of several approaches to investigate the influence of prognostic factors and therapy are given by Ulm *et al.* (1989) and Sauerbrei and Schumacher (1992).

4. Resampling approaches to investigate problems caused by variable selection

If several predictors are available the 'correct' model, which includes all important and no unimportant variables, is unlikely to be selected. Besides other possible model misspecifications, unimportant variables may be included and important variables may be eliminated. In the linear regression model, the elimination of variables with influence on the outcome results

Table 2. Glioma study†

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
(a) Original data set															
Full model															
$\hat{\beta}$	−0.09	−0.06	0.31	0.12	0.45	−0.14	−0.02	−0.31	−0.10	0.04	0.12	−0.13	0.03	0.11	−0.07
SE	0.06	0.07	0.10	0.09	0.08	0.06	0.07	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.06
$\hat{\beta}$/SE	−1.36	−0.91	3.18	1.35	5.75	−2.27	−0.23	−4.90	−1.55	0.69	1.89	−1.80	0.49	1.53	−1.08
BE															
$\hat{\beta}$			0.38		0.43	−0.16		−0.33				−0.14			
SE			0.08		0.07	0.06		0.06				0.07			
$\hat{\beta}$/SE			4.99		5.91	−2.81		−5.70				−2.09			
(b) 100 bootstrap data sets															
Replication															
1	0	0	1	1	1	0	1	1	0	0	1	1	1	0	0
2	1	0	1	0	1	0	0	1	0	0	0	1	1	0	0
3	0	0	1	0	1	1	0	1	1	0	1	1	0	1	1
Inclusion frequency (%)	30	20‡	92	33	100	63	5‡	100	34	11	59	52	20‡	49	22
(c) Selected variables															
Weak factors	0	0	1	0	1	1	0	1	1	0	1	1	0	1	0
Strong factors	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0

†Part (a), parameter estimates $\hat{\beta}$ from the full model and the model selected with BE(0.05) for the standardized values of the original data; part (b) selected models in the first three bootstrap replications and inclusion frequencies over 100 replications using BE(0.05); part (c) variables selected for final models using strategies for weak and for strong factors (1, variable included; 0, variable not included).
‡Negative and positive signs of the corresponding regression parameter estimate.

in biased estimates for the parameter set selected (omission bias), except for the uncommon situation (at least in observational studies) of orthogonality between the selected variables and the non-selected important variables. Even in the orthogonal case the estimates will be biased in logistic models and in models for survival data (Gail *et al.*, 1984). In linear regression models underfitting increases the error sum of squares and the estimate of σ^2. Complex models with many variables may overfit the data. Apart from difficulties in interpreting the contribution of each predictor, overfitting increases the estimates of the variance for each regression estimate and increases the variance of the prediction. The increase depends on the multicollinearity. A penalty term per variable can be used as a parameter to control the complexity of the final model from all-subset selection procedures, and the predefined selection level can be used for stepwise procedures.

4.1. Bootstrap resampling

When the same selection procedure as for the original data is used in an 'ideal' validation study, e.g. on a population defined by the same inclusion and exclusion criteria, then (nearly) the same variables should be selected. This is sometimes called 'replication stability'. Differences between selected variables in two populations may increase when procedures selecting more complex models are used. Harrell *et al.* (1984) used a data splitting approach to investigate the stability of Cox models for survival data chosen by stepwise selection. In three randomly selected subpopulations they found 'great variation in the factors chosen'. Again analysing survival data, Chen and George (1985) took 100 bootstrap samples—as in our approach the patients with complete observations (predictors, outcome and a censoring indicator) were used as the sampling units—and used the same stepwise procedure as the original analysis to select a final model for each bootstrap sample. They could reproduce the original model only in 2% of replications, giving some indication of the instability of the final model. However, the bootstrap inclusion frequencies of five of the six variables from the original model were between 64% and 82%, whereas the frequencies were (much) smaller for variables not included in the original model. This suggested that the most important variables had been selected. Generally, important variables will be included in most of the replications and the inclusion frequencies may be used as a criterion for the importance of a variable. Sauerbrei and Schumacher (1992) extended these stability investigations by considering the inclusion frequencies of all possible pairs of variables to cope with the problem of correlated variables where often only one variable is selected in a bootstrap replication. In an extreme example, one of two highly correlated variables may always be selected, but the inclusion frequency for each variable may be only about 50%. The result may be that both variables are declared unimportant. Sauerbrei and Schumacher (1992) proposed two different variable selection strategies depending on whether only strong factors are of interest or whether weak factors should also be included in the final model. The strategy for weak factors selects in the first step all variables which might be important by keeping a variable in the model if the bootstrap selection frequencies exceed a low level; here we chose 30%. For remaining variables pairwise investigations of interrelationships of inclusion frequencies may show that some of these variables should be eliminated, because their conditional inclusion frequencies are too small if other variables are included. For details see the four steps of strategy A in Sauerbrei and Schumacher (1992). The rationale behind the strategy for strong factors only (strategy B) is as follows: a really strong factor should be entered into the model in nearly all the replications, except when there is another highly correlated variable. In this case at least one of the (two) correlated factors will be selected in every bootstrap replication. Therefore in

the first step all variables with a high selection frequency (here we chose 70%) are included. Variables not included may enter the model in a further step if bidimensional inclusion frequencies suggest that one factor from a 'correlated' pair should be included. Inclusion frequencies of the pair should be larger than 90% and the factor with higher inclusion frequencies is selected for the final model.

4.2. Cross-validation and shrinkage

In his excellent overview of problems introduced by variable selection procedures, Miller (1984) concluded that 'The most important unresolved problem is that of estimation', a statement that is still valid. After having chosen a model, least squares or maximum likelihood estimation of the parameters using the same data set is common practice. This induces 'selection bias'; see Miller (1984) and Miller (1990), pages 110–130. An intuitive explanation was given by Copas and Long (1991):

> 'The choice of the variables to be included depends on estimated regression coefficients rather than their true values, and so X_j is more likely to be included if its regression coefficient is overestimated than if its regression coefficient is underestimated'.

Miller subdivided this selection bias into 'competition bias' and 'stopping rule bias'. The former is the bias introduced by finding the best subset for a fixed number of parameters, whereas the latter is the result of the decision about the criterion for the number of variables (e.g. a simple or a complex model). In assessing the fit of a selected predictor we must distinguish between the retrospective fit (a fit to the same data) and the prospective or validation fit (a fit to new data) (Copas, 1983). The former gives an overoptimistic impression when used to assess the predictive ability for new observations.

Motivated by the predicted residual sum of squares approach, where $\hat{\beta}_{(-i)}$ is used as the parameter estimate of β when the ith observation is eliminated, a (leave-one-out) cross-validated likelihood approach was proposed for generalized linear models and survival time models by van Houwelingen and le Cessie (1990), and extended by Verweij and van Houwelingen (1993). Denoting the log-likelihood by $l(\beta)$, they considered $l_{(-i)}(\beta)$ as the corresponding value with the ith observation eliminated. The contribution of observation i is defined as $l_i(\beta) = l(\beta) - l_{(-i)}(\beta)$. Maximizing $l_{(-i)}(\beta)$ leads to the estimate $\hat{\beta}_{(-i)}$ and they proposed the cross-validated log-likelihood $\text{cvl} = \Sigma\, l_i(\hat{\beta}_{(-i)})$ as a measure of the predictive value of the model.

They also proposed a shrinkage factor based on cross-validation calibration to reduce the bias of parameter estimates caused by model building. For a given model the prognostic index for the ith patient $\text{PI}_i = \mathbf{X}_i\hat{\beta}$, with $\mathbf{X}_i$ as the vector of included variables for patient i and $\hat{\beta}$ the corresponding vector of regression parameter estimates, can be used as the only covariate in a Cox model. Then $\hat{c} = 1$ maximizes the log-likelihood $l(c)$ when c denotes the regression parameter for the prognostic index. Maximizing $l^*(c)$ instead, where l^* is the log-likelihood for the n patients, but using the cross-validated index $\text{PI}_{(-i)} = \mathbf{X}_i\,\hat{\beta}_{(-i)}$ as the only covariate, leads to a regression estimate c^* which is usually smaller than 1, and which was interpreted by Verweij and van Houwelingen (1993) as a shrinkage factor. A value of c^* close to 1 may indicate that hardly any overfitting has occurred, whereas a small value should indicate overfitting. Additionally they proposed $l^*(1)$ as a measure of the fit to new data and therefore as another measure of the predictive ability of a model. The log-likelihood $l^*(1)$ may be used as a model selection criterion.

The global shrinkage factor c^* shrinks each regression coefficient $\hat{\beta}_j$ of an index $\text{PI} = \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$ with the same constant c^* to give for patient i

$$\mathrm{PI}_i(c^*) = c^*(\hat{\beta}_{1(-i)}X_1^i) + \ldots + c^*(\hat{\beta}_{p(-i)}X_p^i).$$

For patient i, X_j^i denotes the value of the included variable X_j. In a simulation study Sauerbrei (1992) showed that selection bias is mainly a severe problem for weak factors, as strong factors will almost always enter a model. The global shrinkage factor may 'overcorrect' and bias towards 0 the effect of a strong variable. Hence we consider, as an extension, PWSFs which can be estimated in the same way by using $\mathrm{PI}_{(-i)}$ and maximizing $l^*(c_1, \ldots, c_p)$. For patient i the prognostic index with shrinkage is then $\mathrm{PI}_i(c_1^*, \ldots, c_p^*) = c_1^*(\hat{\beta}_{1(-i)}X_1^i) + \ldots + c_p^*(\hat{\beta}_{p(-i)}X_p^i)$. Variables whose regression parameter estimates are not biased due to model selection should have a PWSF of about 1. For the estimates of PWSFs in the example data sets, the corresponding predictors were first standardized to mean 0 and standard deviation 1.

5. Results

For the atopy study the results of the full model and the model selected with BE are given in Table 1. For BE we chose a low selection level of 1% because the aim was to develop a model with only 'strong' factors. For all patients 12 of the 23 variables were selected in the final model. This result is in good agreement with the p-values from the Wald test in the full model, where no other variable reached the 5% level. In the full model X_6 and X_7 had a p-value of about 3%, whereas the p-values of all other variables selected are smaller than 0.001. The parameter estimates in the model selected are very similar to those obtained from the full model, but the standard errors are slightly reduced.

Table 1 gives furthermore the parameter estimates and standard errors for the models selected (M_3 and M_4) for two subpopulations defined by the patients whose number ID is even or odd. For each of the subpopulations the number of variables selected is reduced (11 or 8, in contrast with 12 for all patients) and some differences occur in the variables selected. The reduced number of variables selected is caused by the smaller sample size and the corresponding increase in the standard errors. With two exceptions, the estimates of the regression coefficients for selected factors in the subgroup are substantially higher than the corresponding estimates from the models M_1 and M_2 based on all patients.

5.1. Bootstrap resampling

In the glioma study BE(0.05) selects from the original data a final model with five variables. The regression parameter estimates are very similar to the corresponding estimates from the full model, but some of the standardized estimates increase because of a reduction in standard errors (Table 2, part (a)). In the bootstrap approach separate samples were taken from the two treatment groups of the randomized trial to account for the design of the study. Therefore in each replication 205 patients from one treatment arm and 208 patients from the other were included. The results of the first three bootstrap replications and the inclusion frequencies of 100 bootstrap samples are given in part (b) of Table 2. Major differences occur between the models in the different bootstrap data sets and only three variables are selected in nearly all replications. Several variables are selected in about half the replications and some variables are only selected occasionally. For some of these latter variables the signs of the regression parameter estimates in the models selected are positive in some replications and negative in others. With strategy A (section 4.1) of Sauerbrei and Schumacher (1992) for weak factors, a

final model with eight variables is selected. By using BE(0.05) with the original data, five variables were revealed (Table 2), all of which are included in the model for weak factors. Using the strategy for strong factors (strategy B) the variables X_3, X_5 and X_8 were selected because of their high inclusion frequencies. The inclusion of X_6 was dependent on the inclusion of X_{12} and at least one of these two factors was included in 86% of replications. However, neither of these two variables was included in the final model for strong factors, as the inclusion frequency for a pair should be larger than 90%.

Table 3 gives the inclusion frequencies using BE(0.01) for the atopy study. The 'subjective' variables may reduce the diagnostic ability for new patients, and therefore a second score was developed under the objectivity constraint that variables $X_{18}-X_{23}$ were not candidates for the final model. A third score with simplicity for practical use eliminated also the two variables whose values required laboratory measurement. Using strategy B for strong variables (selection level 0.01, all other parameters as in Sauerbrei and Schumacher (1992)) the same variables as in the original analysis were selected for the two models with the constraints of objectivity (model M_5: variables X_3-X_{10} with X_{16} and X_{17}) and simplicity (model M_6: variables X_3-X_{10}). For the unconstrained analysis the variables X_6 and X_7 are not included in the final bootstrap model, and the other variables selected are the same as those in model M_2 for the original data (Table 1). In contrast with the glioma study a cutpoint between important and unimportant variables is apparent. The constraints for objectivity and simplicity further reduce the variability of the variables selected in the bootstrap replications.

Table 3. Atopy study: inclusion frequencies of variables from using the BE method (selection level $\alpha = 0.01$) in 100 bootstrap replications for three analyses with and without constraints on variables allowed in the model†

Variable	*Inclusion frequencies (%) for the following constraints:*		
	None	*Objectivity*	*Simplicity*
sex	3	5	0
age	26	11	15
X_3	97	100	100
X_4	96	99	100
X_5	98	100	100
X_6	50	81	86
X_7	47	84	78
X_8	100	100	100
X_9	100	100	100
X_{10}	100	100	100
X_{11}	9	1	9
X_{12}	22	19	22
X_{13}	2	1	4
X_{14}	7	25	32
X_{15}	20	17	1
X_{16}	33	76	nu
X_{17}	85	71	nu
X_{18}	100	nu	nu
X_{19}	87	nu	nu
X_{20}	98	nu	nu
X_{21}	1	nu	nu
X_{22}	35	nu	nu
X_{23}	2	nu	nu

†nu indicates a variable not used because of a constraint. X_3-X_{15} are objective variables, X_{16} and X_{17} are laboratory measurements and $X_{18}-X_{23}$ are subjective variables.

5.2. Cross-validation and shrinkage

For the glioma study Table 4 gives the number of variables in the model, with the usual log-likelihood $l(1)$ and the log-likelihood $l^*(1)$ based on the corresponding index $PI_{(-i)}$, for the full model and three models selected by using BE. As the models are nested the log-likelihood decreases with the elimination of variables. Although 11 of the 15 variables are eliminated in the model from BE(0.01), the prognostic index is highly correlated with the corresponding prognostic index from the full model (Fig. 1). Pearson's correlation coefficient is 0.94. All other pairwise correlation coefficients between the four indices are higher with the highest being 0.99 (full model and BE(0.157)).

Elimination of the variables does not lead to a decrease in the cross-validated log-likelihood as the values $l^*(1)$ are substantially smaller for the full model than for the models based on BE (Table 4). Using $l^*(1)$ as the criterion, the results indicate that the predictive ability of the most parsimonious model with only four variables seems to be better than that from the full model with 15 variables. The models with five or nine variables appear to have the best predictive ability. The shrinkage factors c^* for the prognostic index based on the models selected with BE(0.05) and BE(0.01) are close to 1, whereas the shrinkage factor for BE(0.157) is smaller and the value 0.825 from the full model seems to indicate a substantial amount of overfitting. The plot of the cross-validated log-likelihoods $l^*(c)$ for the four models (Fig. 2) demonstrates further that the full model has lower values and that the maxima of the likelihoods for the more parsimonious models are closer to 1.

In the model with only four variables none of the PWSFs, shown in Table 4, is substantially smaller than 1. With the exception of X_6 in the full model, the inclusion of additional variables has hardly any effect on the shrinkage factors of these four variables in

Table 4. Glioma study: log-likelihood without (l) and with (l^*) cross-validation for the full model and three models selected by using BE and selection levels 0.157, 0.05 and 0.01—global shrinkage factors and PWSFs for each of the selected models

Variable	*Results for the following models:*			
	Full	*From BE(0.157)*	*From BE(0.05)*	*From BE(0.01)*
Number of variables	15	9	5	4
$l(1)$	−1332.1	−1334.3	−1340.3	−1342.6
$l^*(1)$	−1350.7	−1345.0	−1345.6	−1346.9
Shrinkage factors				
Global	0.825	0.894	0.943	0.952
Parameterwise				
X_1	0.20			
X_2	−0.31			
X_3	1.12	0.97	0.96	0.96
X_4	0.70	0.76		
X_5	0.97	0.98	0.98	0.98
X_6	1.47	0.84	0.84	0.85
X_7	−16.23			
X_8	1.04	0.97	0.94	0.94
X_9	0.74	0.54		
X_{10}	−0.63			
X_{11}	0.77	0.70		
X_{12}	0.90	0.78	0.81	
X_{13}	−1.99			
X_{14}	1.10	0.52		
X_{15}	−0.14			

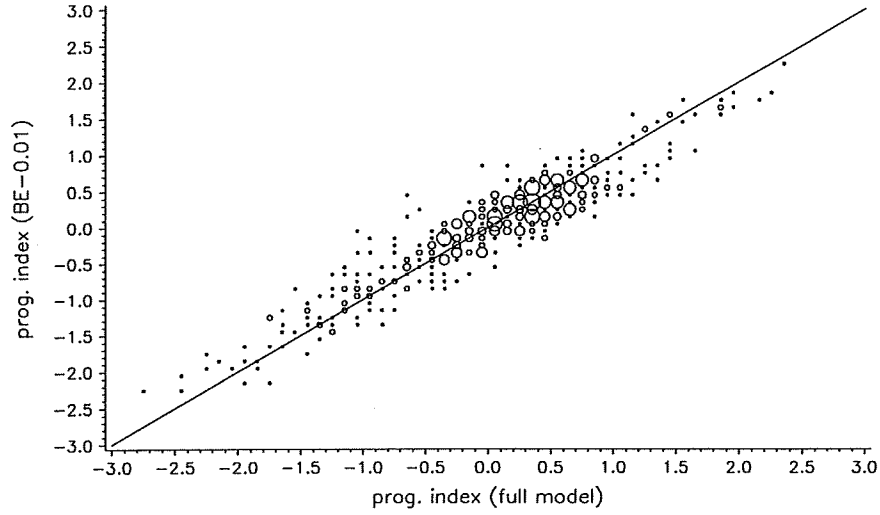

Fig. 1. Scatterplot of the prognostic indices from the full model (15 variables) and from the model selected by BE (selection level 0.01; four variables) for 413 patients in the glioma study: the radius of each circle represents the number of cases

more complex models, but shrinkage factors from other variables which are included in the selected model with BE(0.157) indicate that substantial selection bias is present for X_9 and X_{14}. In the full model some shrinkage factors even have a negative sign indicating that in the original data the sign of the corresponding parameter estimate from the full model may be wrongly estimated. This can be seen as a strong indication that the corresponding variable should not be used in a predictive model. Some of the shrinkage estimates are larger than 1 which may be the result of estimating many, possibly too many, shrinkage factors in this

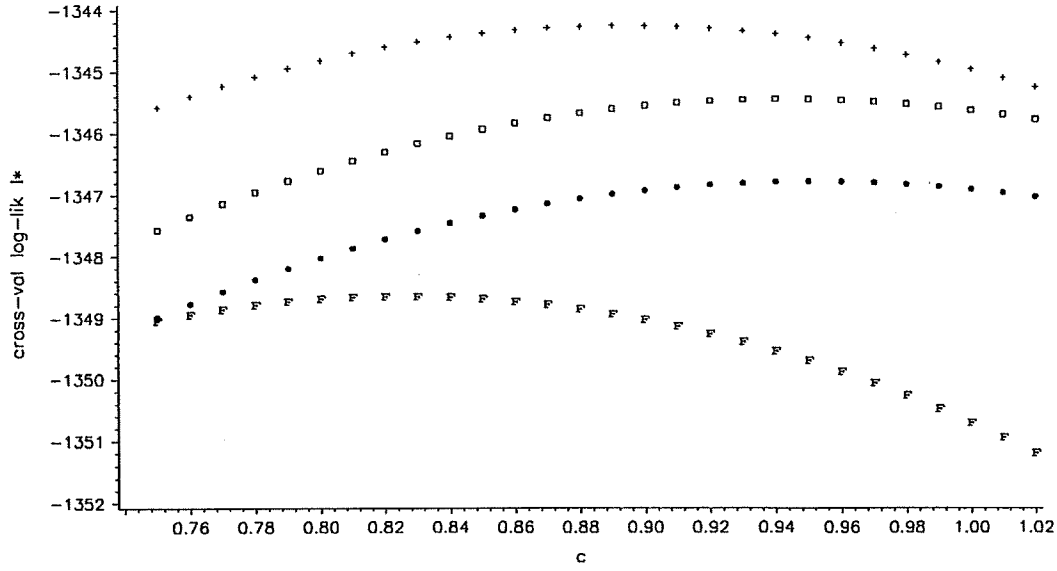

Fig. 2. Cross-validated log-likelihood $l^*(c)$ for the full model (F) and three models selected by BE and selection levels 0.157 (+), 0.05 (□) and 0.01 (•) in the glioma study

model. The surprising value of 1.47 for X_6 in the full model is caused by the inclusion of X_7 with which X_6 is highly correlated, indicating that further investigation is necessary.

The relationship between the PWSFs from the full model and the bootstrap inclusion frequencies may be seen in Tables 2 and 4. The five shrinkage factors with a negative sign correspond to the lowest bootstrap inclusion frequencies (all less than 25%). Variables selected in nearly all the bootstrap replications have estimated shrinkage factors that are close to 1. For the 15 variables Spearman's rank correlation coefficient between the inclusion frequencies and the PWSF is 0.91.

We present the estimated global shrinkage factors and PWSFs for four models derived from the data of the atopy study in Table 5. The global factors decrease with an increase in the number of variables in the model. None of the PWSFs from variables selected by BE is much smaller than 1.

The results from the full model demonstrate empirical evidence for a strong relationship between PWSFs and bootstrap inclusion frequencies. The inclusion frequencies of the five variables with negative shrinkage factors were all less than 10%. Most of the estimated shrinkage factors of the five variables (age, X_{12}, X_{15}, X_{16} and X_{22}) with inclusion frequencies between 20% and 35% were much smaller than 1. Eight of the remaining variables were selected in nearly all the bootstrap replications and have an estimated shrinkage factor of about 1. Only two other variables with a shrinkage factor of about 1 (X_6 and X_7) had inclusion fractions of about 50%. The shrinkage factors of X_{17} and X_{19} with inclusion

Table 5. Atopy study: global shrinkage factors and PWSFs for the full model and three models selected by using BE (selection level 0.01) and various constraints on the inclusions of variables

Variable	*Values for the full model* M_1	*Values for the following constraints:*		
		None, M_2	*Objectivity,* M_5	*Simplicity,* M_6
Global	0.81	0.91	0.94	0.95
Parameterwise				
sex	−3.01			
age	0.75			
X_3	1.20			
X_4	1.24	0.95	0.95	0.96
X_5	1.34	0.90	0.95	0.96
X_6	1.19	0.93	0.95	0.96
X_7	1.16	0.86	0.91	0.94
X_8	1.09	0.76	0.86	0.86
X_9	1.08	0.92	0.94	0.95
X_{10}	1.24	0.91	0.95	0.95
X_{11}	−0.66	0.91	0.95	0.96
X_{12}	0.46			
X_{13}	0.14			
X_{14}	−32.09			
X_{15}	0.27			
X_{16}	1.02			
X_{17}	0.76		0.89	
X_{18}	1.07	0.86	0.91	
X_{19}	1.04	0.90		
X_{20}	1.08	0.88		
X_{21}	−2.65	0.90		
X_{22}	0.86			
X_{23}	−21.73			

frequencies of about 85% are 0.76 and 1.04 respectively. Spearman's rank coefficient between the inclusion frequencies and the PWSFs is 0.84.

6. Discussion

In observational studies at least, statistical models are often determined by data-driven selection methods. These methods usually have only a heuristic basis and their sampling properties are largely unknown. Often many potentially important variables are collected and an essential task of an applied statistician is distinguishing between 'important' and 'unimportant' variables. Additionally the influence of important variables on the outcome must be assessed. At best, some asymptotic results for parts of complex selection procedures are available, but

> 'with a limited set of data and a large set of models asymptotic results for model selection will typically no longer apply' (Hjorth, 1989).

More than 20 years ago data splitting was proposed to investigate some of the problems of variable selection methods, but it suffers the important disadvantage that it reduces the sample size which increases the standard errors and tends to reduce the number of significant variables. In contrast with the analysis presented here, the second part of the data is often reserved for parameter estimation of the final model from the first part. Different approaches which split the data in equal or unequal portions exist; a review on data splitting is given by Picard and Berk (1990). With the development of computer facilities other approaches based on resampling methods are preferable. Using two examples we investigated the complexity and stability of selected models and selection bias of estimated regression parameters.

6.1. Selection level to control complexity of a model

Resampling methods offer new insights into problems caused by complex data-driven model building procedures. Nevertheless, before using a computer for the analysis, a statistician and a scientist should discuss the aim of the study, particularly aspects with consequences for the model building strategy.

Two questions arise.

(a) Is there interest in specific effects of one or more individual variables, or is the only aim a well fitting model with good predictive ability?
(b) Can a complex model with many variables which fits the data very well be useful for prediction, or may it be preferable to have a simpler model with a slightly poorer fit?

Most work on variable selection concentrates on predictive models. Approaches based on all-subsets variable selection usually try to find a compromise between a good fit and a simple model, using a penalty for each variable as part of their selection criteria. As Mallows's C_p and AIC have an asymptotic significance level of about 16% they usually select complex models. This contrasts with BIC whose penalty per variable increases with the sample size and whose selection level is almost always much smaller than AIC or C_p (Teräsvirta and Mellin, 1986). Often many models with different variables have similar apparent predictive ability. This is demonstrated in Fig. 1 for the prognostic index in the glioma study. Searching for the 'best' model among those with a given number of variables adds only a little extra bias compared with the bias introduced by the choice of the number of variables; see also Hjorth (1989). With stepwise procedures a prespecified selection level can be chosen. The simulation

study of Sauerbrei (1992) showed that the actual level is only slightly higher than nominal, so these methods are suitable for controlling the complexity of the model. For a given effect the corresponding standard error and the p-value depend on the sample size. Consequently this approach selects simpler models including only strong factors for small sample sizes and models with more variables in larger studies. The selection level that is chosen should depend on the aim of a study and the given sample size. Whereas the usual 5% level was chosen in the glioma study, the 1% level was preferred for the development of the diagnostic indices in the atopy study (Diepgen *et al.*, 1996). As stated earlier, BE with a selection level 0.157 may be used as a proxy for the all-subset approach with AIC. For the glioma study this approach selected the model with nine variables given in Table 4.

6.2. Bootstrap resampling

The standard analysis of the original data should be complemented by investigations of the stability, of possible biases of regression parameter estimates and of additional effects caused by variable selection. Resampling methods are a useful tool, though for the investigation of complex relationships their theoretical background is not well understood. It must be stressed that they are only heuristic proposals; several alternatives are possible and may be useful. At present simulation studies are promising ways to investigate the possibilities and limitations, because analytical results may be achievable only for parts of the procedures. More formal approaches to formulating a concept called 'model selection uncertainty' in statistical inference were discussed by Hjorth (1994), Chatfield (1995) and Buckland *et al.* (1997). In these approaches resampling methods have a central role.

We showed that bootstrap resampling and cross-validation can offer some insight into the problems caused by variable selection. There are many variations of these methods (e.g. k-fold cross-validation, the parametric bootstrap, the double bootstrap and balanced resampling) and much still remains to be done to explore the possibilities and the limits of the resampling approaches (LePage and Billard, 1992). Using the bootstrap, a statistician will recognize the instability in the model selected, especially when a complex model including several weak factors is chosen. In the glioma example a different model was selected in nearly every replication. There are several explanations for this result. A weak factor has only a low power to enter the final model and will be included in some of the replications. From correlated weak factors usually only one will be selected to represent the corresponding effect of the variables from the 'correlated cluster' of several variables (Sauerbrei and Schumacher, 1992). If the effects are of a similar size the specific 'representative' selected in a bootstrap replication depends on chance. Furthermore variables without influence on the outcome are selected with a probability depending on the selection level. If several are considered in the study, the probability that at least one of them is included in the final model is high. The stability of the selected model decreases with an increase in the number of candidate variables. In the atopy study the use of the lower selection level (0.01) and further constraints which eliminated some of the variables before starting the selection procedure resulted in reduced variation among the models chosen. The importance of this aspect, which implies the incorporation of subject-matter knowledge into initial variable selection, was also stressed by Harrell *et al.* (1996).

The use of model selection for the analysis of studies which are too small leads to a severe bias 'away from zero' for factors with a possibly 'weak' effect, when regression parameter estimation uses the same data set. Sauerbrei and Schumacher (1992) suggested that the bootstrap inclusion frequencies may lead to a more careful interpretation of the importance

of a variable than does the usual standardized parameter estimate. In the glioma example the standardized estimate in the final model from the original data for X_6 is -2.81, indicating a strong effect of that variable (see Table 2), but in the bootstrap replications using BE(0.05) the inclusion rate for this variable was only 63% and the PWSF was 0.84.

In contrast with the analysis here where separate samples were taken from the two treatment groups Sauerbrei and Schumacher (1992) did not consider this design aspect in the bootstrap sampling. The results are very similar and this different way of bootstrapping had no severe effect. That could be expected because the study has a sufficient sample size; however, for small studies the specific bootstrap sampling scheme may have a more severe effect on the result.

In many applications 1000 or even more bootstrap samples are used. That may be necessary if for example bootstrap confidence intervals are required. In our examples we used only 100 bootstrap replications. In an investigation of the stability of the inclusion fractions, Sauerbrei and Schumacher (1992) showed that 100 replications may be sufficient, although they suspected that it may be useful to work with several hundred.

6.3. Cross-validation and shrinkage

The cross-validated likelihood and shrinkage factors can provide additional insight into problems introduced by variable selection and may point to serious overoptimism of the predictive or diagnostic ability of data-dependent models. In the glioma example the cross-validated log-likelihood of the model selected with BE(0.05) was larger than the corresponding log-likelihood from the full model, although 10 of the 15 variables were eliminated. This result is supported by the global shrinkage factors of the full models which were only slightly larger than 0.8 in both examples. The simplest models have a shrinkage factor of about 0.95. For a more detailed discussion of correction for overoptimism based on cross-validation and the bootstrap see van Houwelingen and le Cessie (1990) and Efron and Tibshirani (1997). The proposal of PWSFs suffers from the problem of too many parameters to estimate but the two examples discussed and some further experience seem to demonstrate that some variables should definitely not be included in a predictive model. This can be seen as a strong argument against a full model. Additionally they indicate that in a predictive model only the parameter estimates for weak factors are seriously biased away from 0. This was shown in a simulation study (Sauerbrei, 1992) and was the basis of extending the approach to PWSFs. For the global shrinkage factor we see a disadvantage because it ignores this fact and shrinks all estimates by the same amount. For the three models selected with BE(0.01) in the atopy study, none of the PWSFs was much smaller than 1. This may have resulted from the low selection level (0.01). There are similarities between the approach of variable selection followed by PWSFs and recent proposals which combine subset selection with shrinkage for the development of predictive models (Breiman, 1995; Tibshirani, 1996).

From the results based on cross-validation (Table 4) and the strong correlation between prognostic indices based on simple and complex models (Fig. 1), the models with four or five variables selected with BE(0.01) or BE(0.05) respectively may be preferred as predictive models because they are not too complicated, the parameter estimates seem to be (nearly) unbiased and more complex models are unable to improve the predictive ability substantially.

General guidelines for model building in complex situations which are supported by analytical results or convincing simulation studies have not yet been achieved. Blettner and Sauerbrei (1993) demonstrated in an example that many data-dependent decisions can be necessary and showed that sensible alternative decisions and alternative selection strategies

changed the results. They concluded that details of a final analysis which is highly data driven must be interpreted with caution and that validation of the results with new studies is essential. As this may be difficult, not least because of time, money or even ethical considerations, resampling methods can be used as a substitute, at least for some of the critical aspects.

Acknowledgements

I thank Professor M. Schumacher, Professor M. Blettner and Professor P. Royston for helpful discussions and comments on an earlier draft and Professor Royston for linguistic corrections, Professor D. Krauseneck and Professor T. Diepgen for providing the data, and two referees, the Joint Editor and the Associate Editor for comments which led to a considerable improvement of the presentation.

Readers wishing to obtain the data for specific purposes should contact the author.

References

Blettner, M. and Sauerbrei, W. (1993) The influence of model-building strategies on the results of a case-control study. *Statist. Med.*, **12**, 1325–1338.

Breiman, L. (1992) The little bootstrap and other methods for dimensionality selection in regression: *X*-fixed prediction error. *J. Am. Statist. Ass.*, **87**, 738–754.

——(1995) Better subset regression using the nonnegative garotte. *Technometrics*, **37**, 373–384.

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

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference (with discussion). *J. R. Statist. Soc.* A, **158**, 419–466.

Chen, C.-H. and George, S. L. (1985) The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statist. Med.*, **4**, 39–46.

Copas, J. B. (1983) Regression, prediction and shrinkage (with discussion). *J. R. Statist. Soc.* B, **45**, 311–354.

Copas, J. B. and Long, T. (1991) Estimating the residual variance in orthogonal regression with variable selection. *Statistician*, **40**, 51–59.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Diepgen, T. L., Sauerbrei, W. and Fartasch, M. (1996) Development and validation of diagnostic scores for atopic dermatitis incorporating criteria of data quality and practical usefulness. *J. Clin. Epidem.*, **49**, 1031–1038.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. London: Chapman and Hall.

——(1997) Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Statist. Ass.*, **92**, 548–560.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431–444.

Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. (1984) Regression modeling strategies for improved prognostic prediction. *Statist. Med.*, **3**, 143–152.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statist. Med.*, **15**, 361–387.

Hjorth, U. (1989) On model selection in the computer age. *J. Statist. Planng Inf.*, **23**, 101–115.

——(1994) *Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap*. London: Chapman and Hall.

van Houwelingen, J. C. and le Cessie, S. (1990) Predictive value of statistical models. *Statist. Med.*, **9**, 1303–1325.

LePage, R. and Billard, L. (1992) *Exploring the Limits of Bootstrap*. New York: Wiley.

Miller, A. J. (1984) Selection of subsets of regression variables (with discussion). *J. R. Statist. Soc.* A, **147**, 389–425.

——(1990) *Subset Selection in Regression*. London: Chapman and Hall.

Picard, R. R. and Berk, K. N. (1990) Data splitting. *Am. Statistn*, **44**, 140–147.

Roecker, E. B. (1991) Prediction error and its estimation for subset-selected models. *Technometrics*, **33**, 459–468.

Sauerbrei, W. (1992) Variablenselektion in Regressionsmodellen unter besonderer Berücksichtigung medizinischer Fragestellungen (Variable selection in regression models with application in medical research). *Dissertation*. University of Dortmund, Dortmund.

——(1993) Comparison of variable selection procedures in regression models—a simulation study and practical examples. In *Europäische Perspektiven der Medizinischen Informatik, Biometrie und Epidemiologie* (eds J. Michaelis, G. Hommel and S. Wellek), pp. 108–113. Munich: MMV Medizin.

Sauerbrei, W. and Schumacher, M. (1992) A bootstrap resampling procedure for model building: application to the Cox regression model. *Statist. Med.*, **11**, 2093–2109.

Teräsvirta, T. and Mellin, I. (1986) Model selection criteria and model selection tests in regression models. *Scand. J. Statist.*, **13**, 159–171.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Ulm, K., Schmoor, C., Sauerbrei, W., Kemmler, G., Aydemir, Ü., Müller, B. and Schumacher, M. (1989) Strategien zur Auswertung einer Therapiestudie mit der Überlebenszeit als Zielkriterium. *Biometr. Inform. Med. Biol.*, **20**, 171–205.

Verweij, P. J. M. and van Houwelingen, J. C. (1993) Cross-validation in survival analysis. *Statist. Med.*, **12**, 2305–2314.

Biochem. J. (2009) 422, 313–320 (Printed in Great Britain) doi:10.1042/BJ20090176

online data

Rescue of anaemia and autoimmune responses in *SOD1*-deficient mice by transgenic expression of human *SOD1* in erythrocytes

Yoshihito IUCHI*, Futoshi OKADA*, Rina TAKAMIYA†, Noriko KIBE*, Satoshi TSUNODA*, Osamu NAKAJIMA‡, Kazuyo TOYODA§, Ritsuko NAGAE§, Makoto SUEMATSU†, Tomoyoshi SOGA‖, Koji UCHIDA§ and Junichi FUJII*[1]

*Department of Biochemistry and Molecular Biology, Graduate School of Medical Science, Yamagata University, 2-2-2 Iidanishi, Yamagata 990-9585, Japan, †Department of Biochemistry and Integrative Medical Biology, School of Medicine, Keio University, Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan, ‡Research Laboratory for Molecular Genetics, Yamagata University, 2-2-2 Iidanishi, Yamagata 990-9585, Japan, §Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya 464-8601, Japan, and ‖Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan.

Oxidative stress has been implicated as a cause of various diseases such as anaemia. We found that the *SOD1* [Cu,Zn-SOD (superoxide dismutase)] gene deficiency causes anaemia, the production of autoantibodies against RBCs (red blood cells) and renal damage. In the present study, to further understand the role of oxidative stress in the autoimmune response triggered by *SOD1* deficiency, we generated mice that had the *hSOD1* (human *SOD1*) transgene under regulation of the *GATA-1* promoter, and bred the transgene onto the $SOD1^{-/-}$ background ($SOD1^{-/-}$;$hSOD1^{tg/+}$). The lifespan of RBCs, levels of intracellular reactive oxygen species, and RBC content in $SOD1^{-/-}$;$hSOD1^{tg/+}$ mice, were approximately equivalent to those of $SOD1^{+/+}$ mice. The production of antibodies against lipid peroxidation products, 4-hydroxy-2-nonenal and acrolein, as well as autoantibodies against RBCs and carbonic anhydrase II were elevated in the $SOD1^{-/-}$ mice, but were suppressed in the $SOD1^{-/-}$;$hSOD1^{tg/+}$ mice. Renal function, as judged by blood urea nitrogen, was improved in the transgenic mice. These results rule out the involvement of a defective immune system in the autoimmune response of *SOD1*-deficient mice, because $SOD1^{-/-}$;$hSOD1^{tg/+}$ mice carry the hSOD1 protein only in RBCs. Metabolomic analysis indicated a shift in glucose metabolism to the pentose phosphate pathway and a decrease in the energy charge potential of RBCs in *SOD1*-deficient mice. We conclude that the increase in reactive oxygen species due to *SOD1* deficiency accelerates RBC destruction by affecting carbon metabolism and increasing oxidative modification of lipids and proteins. The resulting oxidation products are antigenic and, consequently, trigger autoantibody production, leading to autoimmune responses.

Key words: autoantibody, oxidative stress, superoxide dismutase, transgenic rescue.

INTRODUCTION

Oxidative stress is intimately involved in aging and ischaemic diseases as well as many other diseases [1]. While excess production of ROS (reactive oxygen species) causes tissue damage, local production of ROS, such as hydrogen peroxide, in limited amounts plays a role in intracellular signalling in response to various extracellular stimuli [2]. SOD (superoxide dismutase) is thought to play a central role in antioxidative systems because of its ability to scavenge superoxide anions, the primary ROS generated from molecular oxygen in cells [3]. Mice that lack *SOD1*, encoding Cu,ZnSOD present in cytoplasm and the intermembrane space of mitochondria [4], show relatively mild phenotypes and grow normally [5,6] compared with mice lacking *SOD2*, encoding MnSOD present in the mitochondrial matrix [7]. *SOD2*-deficient mice die due to dilated cardiomyopathy during the neonatal stage. However, acceleration of the pathological condition induced by interventions, such as ischaemia and reperfusion, in many organs of $SOD1^{-/-}$ mice, is more severe than in organs of $SOD1^{+/+}$ mice [8–10].

Recently, we reported both a marked elevation in the ROS of RBCs (red blood cells) and anaemia in $SOD1^{-/-}$ mice [11]. The anaemic phenotype in $SOD1^{-/-}$ mice was further confirmed by two groups [12,13]. Starzyński et al. [13] reported changes in hepatic iron metabolism in addition to haemolytic anaemia. Severe oxidative damage occurs in RBCs because RBCs carry only Cu,ZnSOD as a superoxide-scavenging enzyme due to a lack of mitochondria. RBCs that are hyperoxic bind oxygen in the lungs (~21 %), and release oxygen in peripheral tissues, which are relatively hypoxic (~2 %). Thus, RBCs undergo cyclic exposure to hyperoxic and hypoxic environments, generating large amounts of superoxide via autoxidation of haemoglobin, which is present in RBCs at a concentration of 5 mM [14]. Reportedly, approx. 3 % of haemoglobin undergoes oxidation and releases superoxide every day *in vivo* [14]. Cu,ZnSOD dismutates superoxide to hydrogen peroxide, which is then reduced to water by glutathione peroxidase, catalase or peroxiredoxin. Cu,ZnSOD-deficient RBCs cannot effectively detoxify superoxide, which itself is not very toxic but is converted into highly deleterious ROS. Peroxynitrite is formed by the reaction of superoxide with NO, which is produced by nitric oxide synthase and is in abundant supply in the blood [15]. Although the rate constant of the reaction of superoxide with NO is greater than with SOD ($\sim 1 \times 10^{10}\ M^{-1} \cdot s^{-1}$ compared with $1–2 \times 10^{9}\ M^{-1} \cdot s^{-1}$), high concentrations of Cu,ZnSOD present in RBCs normally suppress peroxynitrite formation [15]. Since superoxide content remains high in the RBCs of $SOD1^{-/-}$

Abbreviations used: ACR, acrolein; AIHA, autoimmune haemolytic anaemia; APF, 2- [6-(4′-amino)phenoxy-3H-xanthen-3-on-9-yl]benzoic acid; BUN, blood urea nitrogen; CAII, carbonic anhydrase II; CE-MS, capillary electrophoresis MS; DHR, dihydrorhodamine 123; G6PDH, glucose-6-phosphate dehydrogenase; HNE, 4-hydroxy-2-nonenal; HRP, horseradish peroxidase; KO/TgA mice, $SOD1^{-/-}$;$hSOD1^{tgA/+}$ mice; MetHb, methaemoglobin; NZB, New Zealand Black; RBC, red blood cell; ROS, reactive oxygen species; SIN-1, 3-morpholinosydnonimine; SLE, systemic lupus erythaematodes; SOD, superoxide dismutase; hSOD, human SOD; *hSOD1*-Tg, transgenic mice with RBC-specific overexpression of *hSOD1*; TBARS, thiobarbituric acid-reactive substances; TgA mice, $SOD1^{+/+}$;$hSOD1^{tgA/+}$ mice; WST-1, 2-(4-iodophenyl)-3-(4-nitrophenyl)-5-(2,4-disulphophenyl)-2*H*-tetrazolium.

[1] To whom correspondence should be addressed (email jfuji@med.id.yamagata-u.ac.jp).

mice, a large amount of peroxynitrite is formed in the RBC and would oxidize components such as lipids and haemoglobin. Oxidative modification is a known cause of RBC destruction *in vivo* [16] and, hence, would result in anaemia in *SOD1*$^{-/-}$ mice.

In addition, we also reported augmented production of autoantibodies against RBCs in *SOD1*$^{-/-}$ mice, accumulation of immune complexes in glomeruli, and renal dysfunction [11], which are abnormalities typically seen in both AIHA (autoimmune haemolytic anaemia) [17] and SLE (systemic lupus erythaematodes) [18]. In AIHA, it is well documented that the autoantibody-coated RBCs are destroyed by splenic macrophages, but the mechanism that initiates the production of the autoantibodies remains unknown [19]. Based on our findings in *SOD1*-deficient mice, we hypothesized that oxidative modification of components of RBCs may be an underlying mechanism for some autoimmune diseases, such as AIHA and SLE [11].

In the present study, we tried to determine the roles of ROS in the production of autoantibodies and autoimmune responses in *SOD1*$^{-/-}$ mice. Since these mice completely lack *SOD1*, we generated transgenic mice that express *hSOD1* (human *SOD1*) with a *SOD1*-knockout background in an erythroid cell-specific manner. All results obtained thus far are consistent with our hypothesis that oxidative stress triggers an immunogenic response against RBCs by increasing oxidative modification.

EXPERIMENTAL

Animals

SOD1-knockout mice, which have a mixed background of C57BL/6 and b129Sv and were originally established by Matzuk et al. [6], were purchased through Jackson Laboratories (Bar Harbor, ME, U.S.A.), back-crossed more than four times, and bred at our institute. The animal room climate was kept under specific pathogen-free conditions at a constant temperature of 20–22 °C with a 12 h alternating light–dark cycle. The study protocol was approved by the Animal Subjects Committee of the Yamagata University School of Medicine.

CE-MS (capillary electrophoresis MS) analysis

RBCs were isolated from heparinized venous blood samples collected from mice according to the previously reported method [20,21]. Briefly, the samples were centrifuged at 2000 *g* at 4 °C for 10 min, and the cells were washed 3 times and suspended in 10 mM Tris/HCl pH 7.4, to adjust haematocrit values to 15 %. The cell samples were purified by centrifuging at 2000 *g* at 4 °C for 10 min, and the pellets were treated with 0.16 ml of cold methanol containing 300 μM L-methionine sulfone for deproteination. L-Methionine sulfone was also used as the internal standard to validate the recovery or loss of metabolites during sample preparation and CE-MS analysis [22]. Next, 0.16 ml of chloroform and 0.08 ml of distilled water were added and thoroughly mixed. The solution was centrifuged at 12000 *g* at 4 °C for 15 min, and the upper aqueous layer was filtered through a centrifugal filter (Microcone YM-3, Millipore, Tokyo, Japan) to remove proteins. The filtrate was analysed by CE-MS. All CE-MS experiments were performed using an Agilent CE Capillary Electrophoresis System equipped with an air pressure pump, an Agilent 1100 series MSD mass spectrometer, an Agilent 1100 series isocratic high-performance liquid chromatography pump, a G1603A Agilent CE-MS adapter kit, and a G1607A Agilent CE-MS sprayer kit (Agilent Technologies). System control, data acquisition and MSD data evaluation were performed using G2201AA Agilent ChemStation software for CE-MSD.

Generation of transgenic mice expressing hSOD

hSOD1 cDNA was ligated into the NotI site of the vector plasmid, IE3.9int (a gift from Professor Masayuki Yamamoto, Department of Medical Biochemistry, Tohoku University, Sendai, Japan), which has a *GATA-1* promoter [23]. Transgenic mice with RBC-specific overexpression of *hSOD1*, *hSOD1*-Tg, were created at Yamagata University by standard techniques. Fertilized eggs were prepared from BDF1 parents and were microinjected with the recombinant plasmid.

PCR analyses of mouse SOD1 allele and hSOD1 transgene

Mouse *SOD1* genotyping was performed by PCR with specific primers to amplify either the wild-type or the knockout allele, using a mixture of primers. PCR was performed for 40 cycles (40 s at 95 °C, 30 s at 62 °C, 50 s at 72 °C) using AmpliTaq Gold DNA Polymerase (Applied Biosystems, CA, U.S.A.) in a DNA thermal cycler (Bio-Rad, Tokyo, Japan).

Blood testing

RBC content was measured by an automated haematology analyser (Celltac-α Nihon Kohden, Tokyo, Japan). Levels of BUN (blood urea nitrogen) in blood plasma were determined using Fuji Dry-chem 3500V and Fuji Dry-chem slides (Fuji film, Tokyo, Japan).

RBC turnover assay

The RBC life span was assayed by *in vivo* biotinylation, followed by FACS analyses as described previously [11]. RBCs were labelled with biotin *in vivo* by an intravenous injection of 200 μl of a 10 mg/ml solution of *N*-hydroxy succinimide-LC-biotin (Pierce) in PBS (pH 7.4). The percentage of biotinylated cells was calculated as a ratio of positive cells to all RBCs.

Flow cytometry of ROS and IgG bound to RBCs

RBCs (5×10^7 in 200 μl of reaction mixture) were incubated with 25 μM DHR (dihydrorhodamine 123) (Molecular Probes) for 15 min or 10 μM APF {2-[6-(4′-amino)phenoxy-3H-xanthen-3-on-9-yl]benzoic acid} (Sekisui Medical, Tokyo, Japan) for 10 min and washed with PBS three times. The IgG bound to RBCs was assessed for RBCs that had been washed with PBS three times followed by incubation with FITC-conjugated rabbit anti-mouse IgG (100-fold dilution; Dako, Kyoto, Japan) [11]. Fluorescence intensity of Rhodamine 123, APF and FITC bound to RBCs was measured using FACS (FACSCalibur, BD Biosciences, Tokyo, Japan).

Immunoblot analyses

RBCs collected from mice were washed three times with PBS, and lysed in 20 mM Tris/HCl, pH 7.4. The lysate was centrifuged at 17000 *g* for 10 min in a microcentrifuge. Protein concentrations of the supernatant were determined using a BCA (bicinchoninic acid) kit (Pierce). To verify expression of hSOD specifically in erythroid cells, proteins were extracted from seven organs of mice after the RBCs were washed out by circulating PBS. Total proteins (30 μg) were separated on SDS/15 % PAGE and electroblotted onto PVDF membranes (Amersham). After blocking, the blots were incubated with either polyclonal, anti-human Cu,ZnSOD antiserum [24] or polyclonal rabbit anti-nitrotyrosine antibody (Millipore-Upstate, Billerica, MA, U.S.A.) and monoclonal mouse anti-nitrotyrosine antibody (a gift from Professor J. S. Beckman, Environmental Health Sciences Center,