# Variable Selection for Logistic Regression using a Prediction Focussed Information Criterion

**Gerda Claeskens, Christophe Croux and Johan Van Kerckhoven**

ORSTAT, K.U. Leuven,

Naamsestraat 69, B-3000 Leuven, Belgium.

*email:* {gerda.claeskens; christophe.croux; johan.vankerckhoven}@econ.kuleuven.be.

**Abstract**

In biostatistical practice, it is common to use information criteria as a guide for model selection. We propose new versions of the Focussed Information Criterion (FIC) for variable selection in logistic regression. The FIC gives, depending on the quantity to be estimated, possibly different sets of selected variables. The standard version of the FIC measures the Mean Squared Error (MSE) of the estimator of the quantity of interest in the selected model. In this paper we propose more general versions of the FIC, allowing other risk measures such as one based on $L_p$-error. When prediction of an event is important, as is often the case in medical applications, we construct an FIC using the error rate as a natural risk measure. The advantages of using an information criterion which depends on both the quantity of interest and the selected risk measure are illustrated by means of a simulation study and application to a study on diabetic retinopathy.

KEYWORDS: Error rate, Focussed information criterion, Forward selection, Logistic regression, Model selection, Risk measures.

# 1 Introduction

Most clinical trials result in rich datasets with numerous variables of potential influence. Model selection methods are therefore becoming an essential tool for any data analyst. For

an overview of model selection literature, see Burnham and Anderson (2002), George (2000), Spiegelhalter, Best, Carlin and van der Linde (2002) or Claeskens and Hjort (2003). In the Wisconsin Epidemiologic Study of Diabatic Retinopathy (WESDR), for example, (Klein et al, 1984) there are eleven continuous covariates, amongst which are the duration of diabetes and the body mass index, and four binary explicative variables, such as the patient's gender, and the type of his/her area of residence. It is unlikely that all of these variables are important for all uses of the data. Outcome of interest in this study is the presence of retinopathy of any degree and we are in particular interested in the prediction of this event.

Traditional model selection methods such as AIC (Akaike, 1974) or BIC (Schwarz, 1978) select one subset of the covariates, no matter which use of the data will follow. The FIC, focussed information criterion (Claeskens and Hjort, 2003), on the other hand, is developed to select a set of variables which is best for a given focus. Hand and Vinciotti (2003) state that "in general, it is necessary to take the prospective use of the model into account when building it", and address explicitly the prediction problem. Given a patient's specific covariate information, the FIC selects a model that is best for, for example, predicting the presence of the disease of this particular patient. It might happen that one model is good for all patients, however, in the analysis of the WESDR we find different models for different patient groups. In particular, it turns out that the glycosylated hemoglobin level is more important, from a predictive point of view, for patients (both men and women) on a high-level insulin treatment than for patients on a low-level insulin treatment.

The FIC in its original format interprets 'best' model in the sense of minimizing the mean squared error (MSE) of the estimator of the quantity of interest. A novel aspect of this paper is that we introduce focussed model selection based on different risk measures, and not only based on MSE. Especially in the context of prediction of an event, we propose and develop a new focussed information criterion based on the error rate as a risk measure.

In Section 3, we define this FIC based on minimizing the error rate, and give explicit formulae to compute it (see Section 3.1). In addition, we define a general FIC based on $L_p$-loss, and provide expressions for the most commonly used cases, in particular for the mean absolute error (MAE) for $p = 1$. For $p = 2$ we are back to the MSE results of Claeskens and Hjort (2003). Section 4 reports on a simulation study to assess the performance of the FIC, as compared to AIC. Section 5 applies the new model selection criteria to the WESDR data and some concluding remarks are made in Section 6.

# 2 Framework and notation

Assume that a set of data $(x_i, y_i)$ is available, where $x_i$ is a covariate vector of length $p + q$, containing the explicative variables which may be continuous or categorical, and $y_i$ is a 0/1 response variable. The data are distributed according to the following model:

$$P(y_i = 1 \mid x_i) = F(x_i^t \beta) \quad \text{for} \quad 1 \leq i \leq n \tag{1}$$

where $F(\cdot)$ is the inverse logit function $F(u) = 1/\{1+\exp(-u)\}$, and $\beta = (\theta^t, \gamma^t)^t$ is the $p+q$-vector of parameters, where $\theta$ consists of the first $p$ parameters, the ones that we certainly wish to be in the selected model, and $\gamma$ holds the last $q$ parameters, the ones that may potentially be included in the chosen model. While the expressions for the model selection criteria derived in this paper are obtained for the logistic regression model, the ideas transfer immediately to other binary regression models.

Naturally, one can choose a complicated model that incorporates all the variables, even though usually only a few of them are significant. However, such a model is not guaranteed to give the best estimates of the quantity of interest. Adding more variables increases the total variability. Another issue with choosing a complex model is its lack of simplicity: medical researchers often prefer simple models, which are easier to interpret. The goal of this paper

is to select a submodel of the logistic regression model (1), and to use that model to predict the value of the response variable for a "new" observation $x_0$.

The notation used in this paper is largely the same as in Claeskens and Hjort (2003), and the necessary quantities for defining the new FICs will be repeated here. In a local misspecification setting, we specify the true value of the parameter vector as $\beta_{\text{true}} = (\theta_{\text{true}}^t, \gamma_0^t + \delta^t/\sqrt{n})^t$, where $n$ is the sample size and $\gamma_0$ is the value of $\gamma$ for the "null model", i.e. the smallest model we consider, containing only the parameter $\theta$. For the model described above, $\gamma_0$ is equal to zero. The *focus* parameter $\mu = \mu(\beta)$ is a function of the model parameters $\beta$. The score at a covariate value $x_0$ in the logistic model is an example of such a focus parameter, where $\mu(\beta) = \beta^t x_0$. The true value of the parameter of interest is then denoted by $\mu_{\text{true}} = \mu(\beta_{\text{true}})$

For the model selection problem there are potentially $2^q$ estimators of $\mu(\beta)$ to consider, one for each subset $S$ of $\{1, \ldots, q\}$. The model indexed by $S$ contains the parameters $\theta$ and those $\gamma_i$ for which $i \in S$. Practical application might rule out some of these subsets a priori. We denote $\gamma_{0,S^c}$ the known vector of "null" values $\gamma_{0,i}$ for $i \in S^c$, the complement of $S$ with respect to $\{1, \ldots, q\}$ and define $\hat{\mu}_S = \mu(\hat{\theta}_S^t, \hat{\gamma}_S^t, \gamma_{0,S^c}^t)$ the maximum likelihood estimator of $\mu$ in the model indexed by $S$.

Let $J_{n,\text{full}}$ be the estimated $(p+q) \times (p+q)$ information matrix of the full model, that is, the model containing $\theta$ and all $\gamma_i$, $(1, \ldots, q)$. We assume that $J_{n,\text{full}}$ is of full rank, and denote its submatrices $J_{n,00}$, $J_{n,01}$, $J_{n,10}$ and $J_{n,11}$, corresponding to the dimensions of $\theta$ and $\gamma$ respectively. Since the model used is a logistic regression model, straightforward calculations show that

$$J_{n,\text{full}} = \frac{1}{n} \sum_{i=1}^{n} p_i(1-p_i) x_i x_i^t,$$

with $p_i = F(x_i^t \beta_{\text{full}})$ the probability associated with observation $i$. For other choices of the inverse link function $F$, a different expression for $J_{n,\text{full}}$ results. Note that $J_{n,\text{full}}$ is consistently estimated by inserting full model estimators. Let $\pi_S$ be a projection matrix

of size $|S| \times q$, which maps $\nu = (\nu_1, \ldots, \nu_q)^t$ to $\nu_S$, the latter consisting of those $\nu_i$ for which $i \in S$. Too few variables in the model (indexed by set $S$) will cause estimators to be biased. Including too many variables, on the other hand, will inflate the variance. Define now $K_n = J_n^{11} = (J_{n,11} - J_{n,10}J_{n,00}^{-1}J_{n,01})^{-1}$ and $K_{n,S} = (\pi_S K_n^{-1} \pi_S^t)^{-1}$. Two other important quantities are the matrix $M_{n,S} = \pi_S^t K_{n,S} \pi_S$ and vector

$$\omega = J_{n,10}J_{n,00}^{-1}\frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma},$$

with the partial derivatives evaluated at the full model. For example, for the particular choice of parameter of interest $\mu(\beta) = \beta^t x_0$, these derivatives are $\frac{\partial \mu}{\partial \theta} = x_{0,0}$ and $\frac{\partial \mu}{\partial \gamma} = x_{0,1}$, where $x_0$ is partitioned according to $\theta$ and $\gamma$. Finally, define

$$D_n = \hat{\delta}_{\text{full}} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0) \xrightarrow{d} D \sim \mathcal{N}_q(\delta, K_n) \tag{2}$$

(see Hjort & Claeskens (2003) for details and more discussion). Then the maximum likelihood estimator of $\mu$ in the model $S$ has the following limiting distribution (Hjort & Claeskens, 2003, Lemma 3.3)

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S = \left(\frac{\partial \mu}{\partial \theta}\right)^t J_{n,00}^{-1} M + \omega^t(\delta - M_{n,S}K_n^{-1}D), \tag{3}$$

where $M \sim \mathcal{N}_p(0, J_{00})$ is statistically independent of $D$. It is immediate to verify that this distribution has mean and variance given by

$$\lambda_S = E[\Lambda_S] = \omega^t(I_q - M_{n,S}K_n^{-1})\delta, \tag{4}$$

$$\sigma_S^2 = \text{Var}(\Lambda_S) = \tau_0^2 + \omega^t M_{n,S}\omega, \tag{5}$$

with $\tau_0^2 = (\frac{\partial \mu}{\partial \theta})^t J_{n,00}^{-1}(\frac{\partial \mu}{\partial \theta})$ the variance of $\hat{\mu}_\emptyset$ in the null model, which is independent of $S$. Note that this distribution $\Lambda_S$ is normal, with a non-zero mean due to the local misspecification setting.

The distribution of $\Lambda_S$ in (3) is the key result on which the novel model selection criteria are based. The new FICs involve the mean and variance of the limit distribution of $\Lambda_S$,

given in (4) and (5). The expressions presented above are the theoretical values, assuming the limit experiment is valid. In practice we need to estimate the information matrix of the full model $J_{n,\text{full}}$ and derive the needed components from this estimate. We estimate the vector $\delta$, which measures the distance between the null and true model, by $\hat{\delta}_{\text{full}} = \sqrt{n}\hat{\gamma}_{\text{full}}$ as in (2). This leads, first, to maximum likelihood estimates of $\lambda_S$ and $\sigma_S$, the mean and variance of the distribution $\Lambda_S$, in the model $S$ and, second, to an estimate of the information criterion for the submodel $S$.

# 3 Prediction focussed information criteria

In Section 3.1 we derive the FIC taking as risk measure the error rate associated with the prediction of an event, tailored for logistic regression problems. The selected submodel is thus aimed at minimizing the probability of misclassification of a new observation $x_0$, i.e. the probability of incorrectly predicting the associated 0/1 outcome $y_0$.

In Section 3.2 we derive an expression for the FIC based on the $L_p$-error. We then verify this result with the FIC based on Mean Squared Error (MSE, $p = 2$) as obtained in Claeskens & Hjort (2003), and present the explicit expression for the FIC based on the Minimum Absolute Error (MAE, $p = 1$). The expressions for the FIC based on $L_p$-risk hold in a general setting, but in the subsequent sections they will be applied with the log-odds ratio as the focus parameter: $\mu_{\text{true}} = x_0^t\beta_{\text{true}}$ and $\hat{\mu}_S = x_0^t\hat{\beta}_S$. In other words, the score of an observation to predict is the focus parameter. The selected model is then aimed at minimizing the $L_p$-loss when predicting the true score value.

For every considered submodel, indexed by $S$, the focussed information criterion is computed and denoted by $\text{FIC}_S$. We select that subset $S$ of $\{1, \ldots, q\}$ for which $\text{FIC}_S$ is the smallest, this leads to the FIC-selected model which is indexed by the optimal $S$.

## 3.1 The FIC based on Error Rate

Our aim is to construct a selection criterion with the purpose of selecting the model that has the lowest probability of misclassifying a "new" observation $x_0$, assuming that it has been generated from the same model as the "training" data $\{(x_i, y_i) \mid 1 \leq i \leq n\}$. A natural choice for the risk function here, denoted $r_{\text{ER}}(S)$, is the probability of misclassifying the observation $x_0$. The abbreviation ER stands for Error Rate. Define $y_0$ the true response for an observation with covariates $x_0$ as a realization of the 0/1 random variable $Y_0$ with $P(Y_0 = 1 \mid x_0) = F(x_0^t \beta_{\text{true}})$, and let $\hat{y}_{0,S}$ be the predicted response according to the model defined by $S$. Then,

$$r_{\text{ER}}(S) = P(Y_0 = 1 \text{ and } \hat{y}_{0,S} = 0 \mid x_0) + P(Y_0 = 0 \text{ and } \hat{y}_{0,S} = 1 \mid x_0).$$

Due to independence of $Y_0$ and $\hat{y}_{0,S}$, this expression reduces to

$$r_{\text{ER}}(S) = P(Y_0 = 1 \mid x_0)P(\hat{y}_{0,S} = 0 \mid x_0) + P(Y_0 = 0 \mid x_0)P(\hat{y}_{0,S} = 1 \mid x_0),$$

and hence, using the logistic regression model,

$$r_{\text{ER}}(S) = F(x_0^t \beta_{\text{true}})P(x_0^t \hat{\beta}_S < 0) + \{1 - F(x_0^t \beta_{\text{true}})\}P(x_0^t \hat{\beta}_S > 0).$$

The misclassification rate is only concerned with the sign of the estimated log-odds ratio, not with the actual value itself. As focus parameter we set $\mu_{\text{true}} = x_0^t \beta_{\text{true}}$ and $\hat{\mu}_S = x_0^t \hat{\beta}_S$. We use $\Lambda_S$, the limit distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ as in (3), to approximate

$$P(x_0^t \hat{\beta}_S < 0) = P(\hat{\mu}_S < 0) = P\{\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) < -\sqrt{n}\mu_{\text{true}}\}.$$

by

$$\Phi\left(\frac{-\sqrt{n}\mu_{\text{true}} - \lambda_S}{\sigma_S}\right),$$

with $\lambda_S$ and $\sigma_S^2$ as in (4) and (5), and $\Phi(\cdot)$ the cumulative density function of the standard normal distribution. From this, the following approximation is proposed for the risk function

$$r_{\text{ER}}(S) \approx F(\mu_{\text{true}})\Phi\left(\frac{-\sqrt{n}\mu_{\text{true}} - \lambda_S}{\sigma_S}\right) + \{1 - F(\mu_{\text{true}})\}\Phi\left(\frac{\sqrt{n}\mu_{\text{true}} + \lambda_S}{\sigma_S}\right).$$

This risk measure serves as the basis for the *Focussed Information Criterion* based on *Error Rate*. Inserting the estimators, see Section 2, this leads to the FIC based on error rate:

$$\mathrm{FIC}_{\mathrm{ER}}(S) = F(\hat{\mu}_{\mathrm{full}})\Phi\left(\frac{-\sqrt{n}\hat{\mu}_{\mathrm{full}} - \hat{\lambda}_S}{\hat{\sigma}_S}\right) + \{1 - F(\hat{\mu}_{\mathrm{full}})\}\Phi\left(\frac{\sqrt{n}\hat{\mu}_{\mathrm{full}} + \hat{\lambda}_S}{\hat{\sigma}_S}\right),$$

where we estimated $\mu_{\mathrm{true}}$ by $\hat{\mu}_{\mathrm{full}} = \mu(\hat{\beta}_{\mathrm{full}})$. Note that this criterion depends on the value of the covariate vector $x_0$ of the observation to predict. This dependence enters through the focus parameter $\mu$, which is also present in the estimated values of $\lambda_S$ and $\sigma_S$, see (4) and (5).

## 3.2 The FIC based on $L_p$-error

Based on the limit distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\mathrm{true}})$ in equation (3), we derive the expressions for the $L_p$-error of $\hat{\mu}_S$, and this for any subset $S$ of $\{1, \ldots, q\}$ and for any positive $p \geq 1$. This $L_p$-risk measure is defined as the $p^{\mathrm{th}}$ order absolute moment of the limit distribution $\Lambda_S$, $r_p(S) = E(|\Lambda_S|^p)$. After some computations, details of which can be found in the Appendix, the following explicit expression is obtained for integer values of $p$:

$$r_p(S) = \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{p} \binom{p}{j} \sigma_S^j \lambda_S^{p-j} \left\{ \int_{-\frac{\lambda_S}{\sigma_S}}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}}\, \mathrm{d}z + (-1)^p \int_{-\infty}^{-\frac{\lambda_S}{\sigma_S}} z^j \mathrm{e}^{-\frac{z^2}{2}}\, \mathrm{d}z \right\}. \tag{6}$$

This expression can be simplified further, and we find that, for $p$ even,

$$r_p(S) = \frac{1}{\sqrt{\pi}} \sum_{j=0}^{p/2} \binom{p}{2j} 2^j \sigma_S^{2j} \lambda_S^{2n-2j} \Gamma\left(j + \frac{1}{2}\right), \tag{7}$$

while for $p$ odd,

$$\begin{aligned} r_p(S) \;=\; & \frac{1}{\sqrt{\pi}} \sum_{j=0}^{(p-1)/2} \binom{p}{2j} \sigma_S^{2j} |\lambda_S|^{p-2j} 2^j \Gamma\left(j + \frac{1}{2}\right) \\ & + \frac{1}{\sqrt{\pi}} \sum_{j=0}^{p} \binom{p}{j} \sigma_S^j (-|\lambda_S|)^{p-j} 2^{j/2} \Gamma\left(\frac{j+1}{2}, \frac{\lambda_S^2}{2\sigma_S^2}\right). \end{aligned} \tag{8}$$

No such explicit form exist for noninteger values of $p$. We denoted $\Gamma(x)$ for the gamma function evaluated in $x$, and $\Gamma(x, a)$ (for $a > 0$) for the incomplete gamma function. We

point out the dependence of $r_p(S)$ on the focus parameter $\mu$. Different choices of $\mu$ lead to different formulae for the focussed criterion, and as a consequence, may lead to different selected models.

We now give details on two special cases of the FIC based on $L_p$-error. The first case is $\text{FIC}_2$ based on the $L_2$-error, better known as the mean squared error. Henceforth this is denoted as $\text{FIC}_{\text{MSE}}$. This model selection criterion has been extensively discussed in Claeskens and Hjort (2003). We here show that $\text{FIC}_{\text{MSE}}$ is a special case of the general formula in the previous section. From (7), it is easy to see that for $p = 2$,

$$r_2(S) = \frac{1}{\sqrt{\pi}} \left\{ \lambda_S^2 \Gamma\left(\frac{1}{2}\right) + 2\sigma_S^2 \Gamma\left(\frac{3}{2}\right) \right\} = \lambda_S^2 + \sigma_S^2.$$

Applying equations (4) and (5), this is written as

$$r_2(S) = \omega^t (I_q - M_{n,S} K_n^{-1}) \delta\delta^t (I_q - K_n^{-1} M_{n,S}) \omega + {\tau_0}^2 + \omega^t M_{n,S} \omega, \tag{9}$$

which is, up to a constant term, equal to the limit FIC as defined in Claeskens and Hjort (2003). Note that an asymptotically unbiased estimate of $\delta\delta^t$ in (9) is given by $\hat{\delta}\hat{\delta}^t - K_n$. Inserting unbiased estimators leads to

$$\text{FIC}_{\text{MSE}}(S) = \hat{\omega}^t (I_q - M_{n,S} K_n^{-1}) \hat{\delta}\hat{\delta}^t (I_q - K_n^{-1} M_{n,S}) \hat{\omega} + 2\hat{\omega}^t M_{n,S} \hat{\omega}.$$

The other special case that we study is $p = 1$, which leads to a "new" criterion minimizing the mean absolute error, MAE. Equation (8) yields

$$r_1(S) = |\lambda_S| + \frac{1}{\sqrt{\pi}} \left\{ -|\lambda_S| \Gamma\left(\frac{1}{2}, \frac{\lambda_S^2}{2\sigma_S^2}\right) + \sqrt{2}\sigma_S \Gamma\left(1, \frac{\lambda_S^2}{2\sigma_S^2}\right) \right\}.$$

Working out this equation further, we define the Focussed Information Criterion based on MAE as the following consistent estimator of $r_1(S)$

$$\text{FIC}_{\text{MAE}}(S) = 2\hat{\lambda}_S \left\{ \Phi(\frac{\hat{\lambda}_S}{\hat{\sigma}_S}) - \frac{1}{2} \right\} + 2\hat{\sigma}_S \phi(\frac{\hat{\lambda}_S}{\hat{\sigma}_S}),$$

where $\Phi(\cdot)$ is the cumulative distribution function and $\phi(\cdot)$ the density function of the standard normal.

# 4 Simulation study

In this section, a simulation study is presented to examine how well the proposed Focussed selection criteria perform with respect to a benchmark criterion, the Akaike Information Criterion (AIC). In Section 4.1, the particulars of the simulation sampling scheme are detailed. In Section 4.2 we additionally address the issue of model averaging. The results of the simulation are presented in Section 4.3.

## 4.1 Simulation settings

For the simulation study, $n_{\text{test}} = 500$ observations $x_{0,i}$ are independently generated from a normal $\mathcal{N}_5(0, \frac{1}{4}I_5)$ distribution, with $I_5$ the $5 \times 5$ identity matrix. These observations constitute the test sample and remain the same throughout the entire simulation. Then, for each of the $M = 1000$ simulations in the experiment, a training sample of $n_{\text{train}} = 50$ observations $(x_i, y_i)$ is generated, according to the model

$$P(y_i = 1 \mid x_i) = F(\theta + x_i^t \gamma),$$

where $\theta = 0$, $\gamma = (1, -1, 1, -1, 0)^t$ and $x_i \sim \mathcal{N}_5(0, \frac{1}{4}I_5)$. The factor $\frac{1}{4}$ is present so that the generated scores $x_i^t \beta$ are distributed according to a standard normal distribution. For each simulation run, we minimize the information criterion under investigation, hereby forcing the intercept term to be in every model. In this experiment we compare the AIC, $\text{FIC}_{\text{MSE}}$, $\text{FIC}_{\text{MAE}}$ and $\text{FIC}_{\text{ER}}$. In total, $2^5 = 32$ submodels are to be compared, including the "null model" (containing only the intercept $\theta$) and the "full model" (containing $\theta$ and the vector $(\gamma_1, \ldots, \gamma_5)^t$). Within each simulation run, we select one AIC best model, and for each of the $n_{\text{test}}$ observations separately three FIC best models, according to MSE, MAE and ER, respectively. In each of those selected models we estimate the scores by $\hat{\mu}_{0,i} = \hat{\theta} + x_{0,i}^t \hat{\gamma}$, of which its sign determines the predicted value of the corresponding binary $y_{0,i}$ values.

For each separate observation in the test sample, we measure the performance of the model selection criteria via (a) the mean squared error of the predicted score (b) the mean average deviation of the predicted score, and (c) the error rate. The MSE on the predicted score is given by

$$\text{MSE}(\hat{\mu}_{0,i}) = \frac{1}{M}\sum_{j=1}^{M}(\hat{\mu}_{0,i}^{(j)} - \mu_{0,i,\text{true}})^2,$$

with $\hat{\mu}_{0,i}^{(j)}$ the estimated score for validation observation $x_{0,i}$ in simulation run $j$, and $\mu_{0,i,\text{true}}$ the true value of the score. Similarly, the MAE on the predicted score is computed as

$$\text{MAE}(\hat{\mu}_{0,i}) = \frac{1}{M}\sum_{j=1}^{M}|\hat{\mu}_{0,i}^{(j)} - \mu_{0,i,\text{true}}|.$$

The MAE performance measure is sometimes preferred since it is, compared to MSE, less influenced by those simulation runs yielding large deviations from the true values. Finally, the error rate is simulated as

$$\text{ER}_i = \frac{1}{M}\sum_{j=1}^{M} I(\hat{\mu}_{0,i}^{(j)}\mu_{0,i,\text{true}} < 0)$$

where $I(\cdot)$ is the indicator function. If the estimated and the true score have the same sign, they give a zero contribution to the sum in the above $\text{ER}_i$, but if the true and the estimated score yield different values of the corresponding $y_{0,i}$, they contribute to the error rate.

We emphasize that the performance measures are computed for each of the $n_{val}$ observations in the test sample separately. To summarize these $n_{val}$ values, we compute their averages and present a boxplot representation in Figure 1.

## 4.2 Furter particulars

A search across all possible models is only feasible for $q$ relatively small, because the number of possible models to search through increases exponentially with $q$. A forward selection approach is an alternative to an exhaustive search, possibly leading to a different selected

model. Starting from the null model, this iterative procedure adds one variable at a time. Specifically, it adds that variable which yields the lowest value for the information criterion when added to the currently "best" model. This process is repeated until $q+1$ nested models are obtained, ranging from the null model to the full model and indexed by $S_0, S_1, \ldots, S_q$. From these models, we select the model that yields the lowest value for the information criterion.

Model averaging can be applied as an alternative to selecting a single model (see also Hjort & Claeskens (2003)). In this case we construct a weighted average of the estimates in the different models. For each of the nested models obtained during the forward variable selection procedure, we compute this weight as

$$w_j = \frac{\exp\{-\frac{1}{2}\text{xIC}(S_j)\}}{\sum_{k=0}^{q}\exp\{-\frac{1}{2}\text{xIC}(S_k)\}}$$

where $\text{xIC}(S_k)$ is the value of the Information Criterion (AIC, FIC, ...) at the model $S_k$ with $k$ included variables, for $k = 0, \ldots, q$. For each of the submodels $S_j$ a prediction of the score $\mu_0 = x_0^t\beta$ for an observation to be classified is obtained, and these predicted values $\hat{\mu}_{0,S_j}$ then generate the "model-averaged" prediction

$$\hat{\mu}_0 = \sum_{j=0}^{q} w_j \hat{\mu}_{0,S_j}.$$

The advantage of a model averaged estimator is that it has, in general, reduced variability. This will be illustrated in the simulation experiments, where results for the "model-averaged" procedure are reported as well. In the classification literature it is a common strategy to combine several classifiers, see, e.g., Kuncheva (2004) for an overview. Of course, averaging over all possible subsets of the full model, or over any other sequence of models is possible.

All computations are performed using the publicly available software package `R`. In our software we use $\text{AIC}_S = -2\log L(\hat{\beta}_S) + 2(p+|S|)$, with $L(\hat{\beta}_S)$ the likelihood of the estimated model index by $S$, such that lower values indicate better models.

## 4.3 Simulation results

As outlined in Section 4.1, the simulation results in $n_{val} = 500$ values of the MSE, MAE and Error Rate, for prediction based on a submodel selected by AIC, $\text{FIC}_{\text{MSE}}$, $\text{FIC}_{\text{MAE}}$, and $\text{FIC}_{\text{ER}}$. These values are also computed for the model-averaged predictions, discussed in Section 4.2. The boxplots in Figure 1 provide a graphical representation of these 500 values. A log-transformation is applied to the MSE and the MAE, to make their distributions more symmetric. Table 1 complements these plots by giving the averages of the performance measure over the $n_{val} = 500$ values, together with the standard error (SE).

Please insert Figure 1 and Table 1 here.

From Figure 1 it is seen that model averaging significantly improves the performance, at least for the MSE and MAE performance measure. In terms of Error Rate, model averaging does not seem to give an improvement, but neither a worsening of the results obtained with single model selection.

For the Error Rate the results are the most clear cut. From Figure 1, we observe that $\text{FIC}_{\text{ER}}$ performs the best on this criterion, and this remains true if we apply model averaging. So $\text{FIC}_{\text{ER}}$ selects, compared to the other selection criteria, the models which yield the lowest error rates. This should not be too surprising, since the risk measure associated with $\text{FIC}_{\text{ER}}$ is the error rate (to be more precise, the error rate of the limiting experiment), and $\text{FIC}_{\text{ER}}$ selects the model having the smallest value of an approximation of this risk measure. It can be verified that the differences between the average Error Rate of the $\text{FIC}_{\text{ER}}$ is indeed significantly smaller than the other average error rates reported in Table 1, both for single model predictions and for averaged-model predictions (paired comparisons with Tukey correction, P-values $< 0.001$).

While $\text{FIC}_{\text{ER}}$ gives the best results for the Error Rate performance criterion, it performs comparatively much worse for MSE and MAE. But this should not be of much concern, since

if the researcher thinks that another risk measure than Error Rate is more appropriate for his/her prediction problem, he/she should use a variable selection method focussed on that particular risk function.

The two figures on top in Figure 1 show that $\text{FIC}_{\text{MSE}}$ and $\text{FIC}_{\text{MAE}}$ outperform the selection procedure based on AIC when using MSE and MAE as performance criterion. Again, one can show that these differences in average performance are highly significant. After model-averaging, these differences become even more pronounced. This is again as one should expect, since variable selection using $\text{FIC}_{\text{MSE}}$ and $\text{FIC}_{\text{MAE}}$ is aimed at choosing the "best" model as measured by the risks MSE and MAE.

Comparing $\text{FIC}_{\text{MSE}}$ and $\text{FIC}_{\text{MAE}}$ is more difficult. When selecting a single model, the MAE for estimates based on $\text{FIC}_{\text{MAE}}$ is on average slightly worse than for $\text{FIC}_{\text{MAE}}$, although the difference is only minor. In the limiting experiment, such an outcome is not possible, but at the finite-sample level there is no guarantee that the model selected using the $\text{FIC}_{\text{MAE}}$ indeed yields the smallest Mean Absolute Errors. Most important, however, is that, at least in this situation setting, both $\text{FIC}_{\text{MSE}}$ and $\text{FIC}_{\text{MAE}}$ do better than AIC, both for model selection and model averaging.

# 5 Analysis of WESDR Data

In this section we perform model selection for the data of the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) with the methods described in Section 3. The data consists of 691 records of subjects with younger-onset diabetes (the incomplete observations were removed before the analysis). The response variable 'y' is a 0/1 variable where 1 indicates the presence of retinopathy of any degree. The continuous covariates are 'rere' and 'lere', the refractive error in diopters for resp. the right and the left eye; 'reip' and 'leip', the internal eye pressure in mmHg for resp. the right and the left eye; 'adia', the age

in years at which diabetes was diagnosed; 'ddia', the duration of diabetes in years; 'gly', the percentage of glycosylated hemoglobin, 'sysp' and 'diap', the resp. systolic and diastolic blood pressure in mmHg; 'bmi', the Body Mass Index, and 'pulse', the pulse rate in beats per 30 seconds. The binary 0/1 covariates are 'sex', with 1 indicating male; 'uri', with 1 indicating the presence of urine protein; 'ins', with 1 indicating more than 1 dose of insulin taken per day, and 'urb', with 1 indicating that the subject lives in an urban county. We refer to Klein et al. (1984) for further discussion of the variables in this data set.

We examine the predictive power of the models selected by the different selection criteria AIC, $\text{FIC}_{\text{MSE}}$, $\text{FIC}_{\text{MAE}}$, $\text{FIC}_{\text{ER}}$ as well as the model-averaged version by assessing their error rates. Note that, since we work here with real data for which the true value of the scores is not available, the MSE and MAE performance criteria cannot be computed. The error rate is estimated by means of a cross-validation experiment: for each patient in the dataset, we select and estimate a model based on all the other patients in the dataset and then make a prediction for the presence of retinopathy of the left-out observation. The model search includes an intercept to all of the models, and allows inclusion or exclusion of all remaining $q = 15$ variables. Then, we compare the predictions with the real values of 'y', the presence of retinopathy of any degree. We count the percentage of wrong predictions, which yields an estimate of the error rate. The results are summarized in Table 2.

Please insert Table 2 here.

We observe from Table 2 that the models selected by the focussed information criteria and the model-averaged estimates based on FIC, all yield a lower error rate than their AIC counterparts. The McNemar test (e.g. Kuncheva 2004, page 13-15) reveals that this difference is strongly significant ($p$-values $< 0.025$). On the other hand, the difference between the error rates for the models selected by the different FICs is not statistically significant. These results illustrate the advantage of selecting a possibly different set of predictor variables

for every observation to predict. Indeed, there is a priori no reason why a unique selected model would be best for all future predictions to be made.

To illustrate that the model selected by the FIC might depend on the observation, we performed a second analysis. We divided the patients into four groups, according to their gender and the number of doses of insulin taken each day, as shown below.

| Group | characteristics |
|---|---|
| A | females taking none or a single insulin dose each day |
| B | females taking multiple insulin doses each day |
| C | males taking none or a single insulin dose each day |
| D | males taking multiple insulin doses each day |

The groups have roughly an equal number of observations. We record for each group the percentage of times that each variable enters the model when predicting an observation belonging to that group. Table 3 shows the selection frequencies for the four most often selected variables in every group, for $\text{FIC}_{\text{MSE}}$ and $\text{FIC}_{\text{ER}}$.

Please insert Table 3 here.

Both FIC methods select the variable 'ddia' most often, and in particular the error rate based FIC has a strong preference for this variable. A logistic regression model containing only an intercept and this variable 'ddia' performs very well, with a cross-validated error rate of 0.1888. In fact, the model selected using $\text{FIC}_{ER}$ ends up with this simple model in 46.3% of the cases. But, as follows from Table 2, the $\text{FIC}_{ER}$ approach reaches even a lower error rate by deviating from this simple model for an important part of the observations to classify. A possible strategy for a more refined analysis is to include the variable 'ddia' in the list of fixed variables which are included in every selected model, together with the intercept.

The second most selected variable is 'gly', the percentage of glycosylated hemoglobin, which is selected about half of the time by the FIC based on MSE, and with a lower frequency by the FIC based on error rate. Variable selection based on $\text{FIC}_{\text{ER}}$ includes the variable 'gly' much more often for groups B and D than for groups A and C. Hence, glycosylated hemoglobin level is less important, from a predictive point of view, for patients taking none or only a single dose of insulin each day (groups A and C) than for patients taking multiple doses of insulin each day (groups B and D).

# 6 Discussion

In this paper, we extended the focused information criterion, as developed by Claeskens and Hjort (2003). It is originally constructed to select a submodel minimizing the mean squared error of the estimator of the focus point. The idea put forward in this paper is that MSE is not the only risk measure that one can consider. We expand the construction and application to minimize the more general $L_p$-norm, of which MSE ($p = 2$) and mean absolute deviation ($p = 1$) are special cases. Another, perhaps more important, contribution of this paper is the proposal of a Focussed Information Criterion using the error rate as risk measure. This is of specific use in binary regression problems, where the goal is to select models which yield the lowest error rate.

To show the usefulness of these information criteria, we presented both a simulation study and an analysis of the WESDR dataset. In these analyses, we observed that the focussed information criteria select models which perform significantly better, for their specific focus (that is, lower MSE for the FIC based on MSE, and lower error rate for the FIC based on error rate), than the Akaike information criterion. In the WESDR data analysis, it was illustrated how different models are selected for different patients. By allowing the selected model to vary with the observation to predict, we obtained a gain in predictive performance.

The variable selection problem becomes even more pertinent when a large number of variables relative to sample size is available. In this setting, the non-existence of the classical logistic regression estimator may cause problems. It is a topic of our current research to apply model selection methods to such data sets.

**Acknowledgment:** The research of the second author has been supported by the Research Fund K.U. Leuven and the Fund for Scientific Research Flanders (Contract number G.0385.03).

# References

Akaike, H. (1974). A new look at statistical model identification, *I.E.E.E. Transactions on Automatic Control,* **19**, 716–723.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer, New York.

Claeskens G., and Hjort N. L. (2003). The focused information criterion [with discussion]. *Journal of the American Statistical Association*, **98**, 900–916.

George E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, **95**, 1304–1308.

Hand D. J., V. Vinciotti (2003). Local versus global models for classification problems: fitting models where it matters. *The American Statistician*, **57**, 124–131.

Hastie T., Tibshirani R., Friedman J. (2001). *The elements of statistical learning: data mining, inference and prediction.* Springer, New York.

Hjort N. L., and Claeskens G. (2003). Frequentist model average estimators [with discussion]. *Journal of the American Statistical Association*, **98**, 879–899.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1984). The Wisconsin epidemiologic study of diabetic retinopaty: II. Prevalence and risk of diabetic

retinopathy when age at diagnosis is less than 30 years, *Archives of Ophthalmology*, **102**, 520–526.

Kuncheva L. I. (2004). *Combining pattern classifiers: methods and algorithms.* Wiley Interscience.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit [with discussion]. *Journal of the Royal Statistical Society*, **B 64**, 583–639.

# A Appendix

*Computation of the $L_p$-norm related risk $r_p(S)$ in equations* (6), (7), *and* (8).

For $\Lambda_S \sim \mathcal{N}(\lambda, \sigma^2)$, we write $E(|\Lambda_S|^p) = E(|\sigma Z + \lambda|^p)$ where $Z$ has a standard normal distribution. From this it follows:

$$\begin{aligned} E(|\Lambda_S|^p) &= \frac{1}{\sqrt{2\pi}} \int_{-\frac{\lambda}{\sigma}}^{+\infty} (\sigma z + \lambda)^p \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z + (-1)^p \int_{-\infty}^{-\frac{\lambda}{\sigma}} (\sigma z + \lambda)^p \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{p} \binom{p}{j} \sigma^j \lambda^{p-j} \left\{ \int_{-\frac{\lambda}{\sigma}}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z + (-1)^p \int_{-\infty}^{-\frac{\lambda}{\sigma}} z^j \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \right\}. \end{aligned}$$

For $p$ even, say $p = 2r$, the expression can be simplified as follows.

$$\begin{aligned} E[|\Lambda_S|^{2r}] &= \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{2r} \binom{2r}{j} \sigma^j \lambda^{2r-j} \int_{-\infty}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \\ &= \sqrt{\frac{2}{\pi}} \sum_{j'=0}^{r} \binom{2r}{2j'} \sigma^{j'} \lambda^{2r-2j'} \int_{0}^{+\infty} z^{2j'} \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \\ \overset{u=\frac{z^2}{2}}{=} & \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \binom{2r}{2j'} 2^{j'} \sigma^{2j'} \lambda^{2r-2j'} \int_{0}^{+\infty} u^{j'-1/2} \mathrm{e}^{-u} \, \mathrm{d}u \\ &= \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \binom{2r}{2j'} 2^{j'} \sigma^{2j'} \lambda^{2r-2j'} \Gamma\left(j' + \frac{1}{2}\right). \end{aligned}$$

For $p$ odd, say $p = 2r+1$, this leads to

$$
\begin{aligned}
E[|\Lambda_S|^p] &= \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{p} \binom{p}{j} \sigma^j \lambda^{p-j} \left\{ \int_{-\frac{\lambda}{\sigma}}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z - (-1)^j \int_{\frac{\lambda}{\sigma}}^{+\infty} z^j \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \right\} \\
&= \frac{1}{\sqrt{2\pi}} \sum_{j'=0}^{r} \left\{ \begin{array}{l} \binom{2r+1}{2j'} \sigma^{2j'} \lambda^{2r+1-2j'} \left\{ \int_{-\frac{\lambda}{\sigma}}^{+\infty} z^{2j'} \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z - \int_{\frac{\lambda}{\sigma}}^{+\infty} z^{2j'} \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \right\} \\ + \binom{2r+1}{2j'+1} \sigma^{2j'+1} \lambda^{2r-2j'} \left\{ \int_{-\frac{\lambda}{\sigma}}^{+\infty} z^{2j'+1} \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z + \int_{\frac{\lambda}{\sigma}}^{+\infty} z^{2j'+1} \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \right\} \end{array} \right\} \\
&= \sqrt{\frac{2}{\pi}} \sum_{j'=0}^{r} \left\{ \begin{array}{l} \binom{2r+1}{2j'} \sigma^{2j'} \lambda^{2r+1-2j'} \operatorname{sign}(\lambda) \int_0^{\frac{|\lambda|}{\sigma}} z^{2j'} \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \\ + \binom{2r+1}{2j'+1} \sigma^{2j'+1} \lambda^{2r-2j'} \int_{\frac{|\lambda|}{\sigma}}^{+\infty} z^{2j'+1} \mathrm{e}^{-\frac{z^2}{2}} \, \mathrm{d}z \end{array} \right\} \\
\overset{u=\frac{z^2}{2}}{=} \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \left\{ \begin{array}{l} \binom{2r+1}{2j'} \sigma^{2j'} \lambda^{2r+1-2j'} \operatorname{sign}(\lambda) 2^{j'} \int_0^{\frac{\lambda^2}{2\sigma^2}} u^{j'-\frac{1}{2}} \mathrm{e}^{-u} \, \mathrm{d}u \\ + \binom{2r+1}{2j'+1} \sigma^{2j'+1} \lambda^{2r-2j'} 2^{j'+1/2} \int_{\frac{\lambda^2}{2\sigma^2}}^{+\infty} u^j \mathrm{e}^{-u} \, \mathrm{d}u \end{array} \right\} \\
&= \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \left\{ \begin{array}{l} \binom{2r+1}{2j'} \sigma^{2j'} \lambda^{2r+1-2j'} \operatorname{sign}(\lambda) 2^{j'} \{\Gamma(j'+\frac{1}{2}) - \Gamma(j'+\frac{1}{2}, \frac{\lambda^2}{2\sigma^2})\} \\ + \binom{2r+1}{2j'+1} \sigma^{2j'+1} \lambda^{2r-2j'} 2^{j'+1/2} \Gamma(j'+1, \frac{\lambda^2}{2\sigma^2}) \end{array} \right\} \\
&= \frac{1}{\sqrt{\pi}} \sum_{j'=0}^{r} \binom{2r+1}{2j'} \sigma^{2j'} |\lambda|^{2r+1-2j'} 2^{j'} \Gamma\left(j' + \frac{1}{2}\right) \\
&\quad + \frac{1}{\sqrt{\pi}} \sum_{j=0}^{2r+1} \binom{2r+1}{j} \sigma^j (-|\lambda|)^{2r+1-j} 2^{j/2} \Gamma\left(\frac{j+1}{2}, \frac{\lambda^2}{2\sigma^2}\right).
\end{aligned}
$$

This ends the proof.

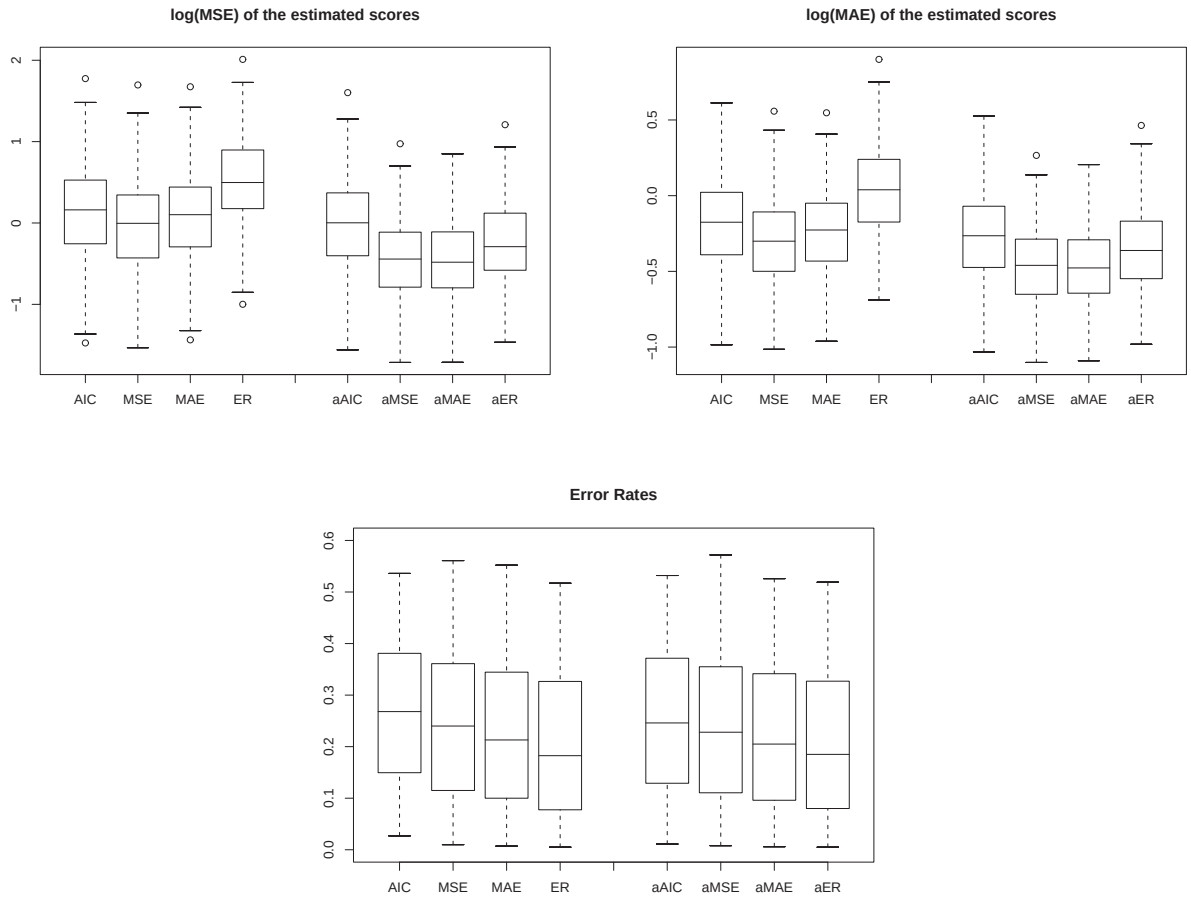


Figure 1: Boxplots of the log(MSE), log(MAE) and Error Rates of the 500 observations in the test sample. The MSE, MAE, and Error rates have been simulated for estimators from a model selected by the criteria AIC, $\text{FIC}_{\text{MSE}}$, $\text{FIC}_{\text{MAE}}$, or $\text{FIC}_{\text{ER}}$, as well as for the model averaged versions of the estimators (indicated by the prefix "a")

Table 1: Average values, together with their standard errors, of the log(MSE), log(MAE) and Error Rates over the 500 observations in the test sample. The MSE, MAE, and Error rates have been simulated for estimators from a model selected by the criteria AIC, $\text{FIC}_{\text{MSE}}$, $\text{FIC}_{\text{MAE}}$, and $\text{FIC}_{\text{ER}}$, as well as for the model averaged versions of the estimators (indicated by the prefix "a") .

| | log(MSE) | | log(MAE) | | Error Rate | |
|---|---|---|---|---|---|---|
| Criterion | Average | SE | Average | SE | Average | SE |
| AIC | 0.14091 | 0.02490 | -0.18210 | 0.01277 | 0.26621 | 0.00599 |
| $\text{FIC}_{\text{MSE}}$ | -0.02613 | 0.02448 | -0.29752 | 0.01269 | 0.24650 | 0.00643 |
| $\text{FIC}_{\text{MAE}}$ | 0.08465 | 0.02390 | -0.23785 | 0.01193 | 0.22875 | 0.00646 |
| $\text{FIC}_{\text{ER}}$ | 0.50670 | 0.02396 | 0.03379 | 0.01337 | 0.20750 | 0.00649 |
| $a$AIC | -0.01428 | 0.02423 | -0.27092 | 0.01235 | 0.25002 | 0.00636 |
| $a\text{FIC}_{\text{MSE}}$ | -0.44985 | 0.02117 | -0.46220 | 0.01107 | 0.23927 | 0.00645 |
| $a\text{FIC}_{\text{MAE}}$ | -0.46620 | 0.02129 | -0.47364 | 0.01075 | 0.22336 | 0.00640 |
| $a\text{FIC}_{\text{ER}}$ | -0.25137 | 0.02259 | -0.35661 | 0.01244 | 0.20913 | 0.00645 |

Table 2: Error rates for the WESDR data, obtained via cross-validation. The models are selected using AIC, $\text{FIC}_{\text{MSE}}$, $\text{FIC}_{\text{MAE}}$ $\text{FIC}_{\text{ER}}$ and also results for the model-averaged estimates are reported (indicated by the prefix "a") .

| Method | AIC | $\text{FIC}_{\text{MSE}}$ | $\text{FIC}_{\text{MAE}}$ | $\text{FIC}_{\text{ER}}$ | $a$AIC | $a\text{FIC}_{\text{MSE}}$ | $a\text{FIC}_{\text{MAE}}$ | $a\text{FIC}_{\text{ER}}$ |
|---|---|---|---|---|---|---|---|---|
| Error Rate | 0.198 | 0.174 | 0.174 | 0.177 | 0.193 | 0.172 | 0.174 | 0.174 |

Table 3: Model selection methods $\text{FIC}_{\text{MSE}}$ and $\text{FIC}_{\text{ER}}$ are applied to each subject within a group of the WESDR data. The table shows the selection percentages of the four most frequently selected variables per group.

| | Group | Variable 1 | | Variable 2 | | Variable 3 | | Variable 4 | |
|---|---|---|---|---|---|---|---|---|---|
| $\text{FIC}_{\text{MSE}}$ | A | ddia | 86.2% | gly | 53.8% | pulse | 42.6% | reip | 39.0% |
| | B | ddia | 81.8% | gly | 50.0% | pulse | 33.8% | urb | 32.4% |
| | C | ddia | 78.5% | gly | 51.3% | pulse | 34.4% | reip | 33.8% |
| | D | ddia | 77.8% | gly | 54.9% | reip | 39.2% | pulse | 37.9% |
| $\text{FIC}_{\text{ER}}$ | A | ddia | 92.3% | gly | 28.2% | reip | 17.4% | uri | 16.9% |
| | B | ddia | 90.5% | gly | 45.3% | uri | 33.8% | diap | 25.0% |
| | C | ddia | 89.2% | gly | 36.4% | uri | 31.8% | bmi | 24.6% |
| | D | ddia | 90.8% | gly | 41.8% | uri | 32.0% | pulse | 28.8% |