Variable Selection for Logistic Regression using a Prediction Focussed Information Criterion

Gerda Claeskens, Christophe Croux and Johan Van Kerckhoven

ORSTAT and University Center for Statistics, 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 Diabetic 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 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 and BIC. 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 $d + q$, 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 $(d + q)$-vector of parameters, where θ consists of the first d parameters, the ones that we certainly wish to be in the selected model, and γ 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 logistic regression models, 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 γ_0 is the value of γ for the "null model", i.e. the smallest model we consider, containing only the parameter θ. For the model described above, γ_0 is equal to zero. The *focus* parameter $\mu = \mu(\beta)$ is a function of the model parameters β. The linear predictor at a covariate value x_0 in the logistic model is an example of such a focus parameter, where $\mu(\beta) = x_0^t\beta$. 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 $\{1, \ldots, q\}$. Other estimation methods, such as model averaging or shrinkage estimators, combine several of these submodel estimators. The model indexed by S contains the parameters θ and those γ_i for which $i \in S$. In practical applications, the user might rule out some of these subsets a priori. We denote γ_{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 $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$ the maximum likelihood estimator of μ in the model indexed by S.

Let $J_{n,\text{full}}$ be the estimated $(d+q) \times (d+q)$ information matrix of the full model, and J_{full} the limiting information matrix. 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 θ and γ respectively, and analogously for J_{full}. Since the model used is a logistic regression model, straightforward calculations show that

$$J_{n,\text{full}} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\beta\partial\beta^t} \log f(Y_i \mid x_i, \beta) = \frac{1}{n}\sum_{i=1}^{n} p_i(1-p_i)x_i x_i^t, \tag{2}$$

with $f(\cdot)$ the binomial probability mass function, and $p_i = F(x_i^t\beta)$ the probability associated

with observation i. For other choices of the inverse link function F, a different expression for $J_{n,\text{full}}$ results. In practice we insert for β in $J_{n,\text{full}}$ the full model estimator.

First define $K = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, the limiting variance of $\hat{\gamma}$ in the full model, and K_n its finite sample counterpart. Then we have

$$D_n = \hat{\delta}_{\text{full}} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0) \xrightarrow{d} D \sim \mathcal{N}_q(\delta, K), \tag{3}$$

where δ measures the distance between the null and true model (see Hjort & Claeskens (2003) for details and more discussion). The maximum likelihood estimator of μ in the model S has now the following limiting distribution (Hjort & Claeskens, 2003, Lemma 3.3)

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S = \left(\frac{\partial \mu}{\partial \theta}\right)^t J_{00}^{-1} M + \omega^t(\delta - M_S K^{-1} D), \tag{4}$$

where $M \sim \mathcal{N}_d(0, J_{00})$ is statistically independent of D. Here we use the quantities $M_S = \pi_S^t(\pi_S K^{-1} \pi_S^t)^{-1}\pi_S$, the limiting variance of $(\hat{\gamma}_S, \gamma_{0,S^c})$, and $M_{n,S}$ its finite sample counterpart, and where π_S stands for the projection matrix of size $|S| \times q$, mapping any vector $\nu = (\nu_1, \ldots, \nu_q)^t$ to ν_S, the latter consisting of those ν_i for which $i \in S$. We also need the auxiliary vector $\omega = J_{10}J_{00}^{-1}\frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$, where we evaluate the partial derivatives at the full model. For example, for the particular choice of parameter of interest $\mu(\beta) = x_0^t\beta$, 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 the dimensions of θ and γ.

Some calculations yield that the limiting distribution Λ_S has mean and variance

$$\lambda_S = E[\Lambda_S] = \omega^t(I_q - M_S K^{-1})\delta, \tag{5}$$

$$\sigma_S^2 = \text{Var}(\Lambda_S) = \tau_0^2 + \omega^t M_S \omega, \tag{6}$$

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

The new FICs involve the mean and variance of the limiting distribution of Λ_S, given in (5) and (6). The expressions presented above are the theoretical values, assuming the limiting 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 δ by $\hat{\delta}_{\text{full}} = \sqrt{n}\hat{\gamma}_{\text{full}}$ as in (3). This leads, first, to maximum likelihood estimators of λ_S and σ_S^2, the mean and variance of the distribution Λ_S, in the model S and, second, to an estimator of the information criterion for the submodel S.

3 Prediction focussed information criteria

The traditional AIC and BIC information criteria are, as FIC, based on a likelihood approach. Where the FIC takes on different values, depending on a specified focus parameter, the AIC or BIC values do not depend on the purpose of the statistical analysis. In this section we show how the results of Claeskens & Hjort (2003) can be applied for obtaining focussed information criteria when prediction of a binary variable is of interest.

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. 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 Mean 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 are applied with the linear predictor of an observation, here the log-odds ratio, as the focus parameter: $\mu_{\text{true}} = x_0^t\beta_{\text{true}}$ and $\hat{\mu}_S = x_0^t\hat{\beta}_S$.

The selected model is then aimed at minimizing the L_p-loss when predicting the true value of the focus parameter.

For every considered submodel, indexed by S, the focussed information criterion is com-

puted and denoted by FIC_S. We select that subset S of $\{1, \ldots, q\}$ for which 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,

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

This misclassification rate is only concerned with the sign of the estimated log-odds ratio, not with the actual value itself. We now apply the methodology of Claeskens & Hjort (2003), with $\mu_{\text{true}} = x_0^t \beta_{\text{true}}$ as focus parameter, and $\hat{\mu}_S = x_0^t \hat{\beta}_S$. We emphasize that our ultimate goal is prediction, rather than parameter estimation, and we only define a focus parameter for mathematical reasons, such that the results of Claeskens & Hjort (2003) can be applied.

We use Λ_S, the limiting distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ as in (4), 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 $\Phi\left\{-(\sqrt{n}\mu_{\text{true}} + \lambda_S)/\sigma_S\right\}$, with λ_S and σ_S^2 as in (5) and (6), and $\Phi(\cdot)$ the cumulative density function of the standard

normal distribution. From this, the following approximation is proposed for the risk function

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

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 μ_{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 through the focus parameter μ, which is also present in the estimated values of λ_S and σ_S, see (5) and (6).

3.2 The FIC based on L_p-error

Based on the limiting distribution of $\sqrt{n}(\hat{\mu}_S - \mu_{\mathrm{true}})$ in equation (4), 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 $p \geq 1$. The L_p-risk measure is defined as the p^{th} order absolute moment of the limiting distribution Λ_S, $r_p(S) = E(|\Lambda_S|^p)$. Note that we work with the absolute moments and not the centered ones because we want a measure of the deviations of $\hat{\mu}_S$ to μ, and the bias involved should not be eliminated by centering. For integer values of p it is possible to derive an explicit expression for $r_p(S)$. The general expressions, and details on their derivation, are in the technical report available at http://www.tibs.org/biometrics. Note again the dependence of $r_p(S)$ on the focus parameter: different choices of μ will 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 FIC_2 based on the L_2-error, better known as the mean squared error and henceforth denoted as $\mathrm{FIC}_{\mathrm{MSE}}$. This model selection criterion has been extensively discussed in Claeskens and

Hjort (2003). For $p = 2$, $r_2(S) = \lambda_S^2 + \sigma_S^2$. Applying equations (5) and (6), this can be written as

$$r_2(S) = \omega^t(I_q - M_{n,S}K_n^{-1})\delta\delta^t(I_q - K_n^{-1}M_{n,S})\omega + {\tau_0}^2 + \omega^t M_{n,S}\omega, \tag{7}$$

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 (7) is given by $\hat{\delta}\hat{\delta}^t - K_n$. Inserting unbiased estimators leads to

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

The other special case that we study is $p = 1$, which leads to a "new" criterion minimizing the mean absolute error, MAE. Here it can be verified that

$$r_1(S) = 2\lambda_S\left\{\Phi(\frac{\lambda_S}{\sigma_S}) - \frac{1}{2}\right\} + 2\sigma_S\phi(\frac{\lambda_S}{\sigma_S}).$$

Then we define the Focussed Information Criterion based on MAE as the following estimator of $r_1(S)$

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

where $\phi(\cdot)$ is the 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 two better known criteria, the Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). 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 $\mathcal{N}_q(0, \frac{1}{4}I_q)$ distribution, with I_q the $q \times q$ 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 size n_{train} 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 $\theta = 0$, $\gamma = (1, -1, 1, -1, 0, \ldots, 0)^t$ such that only 4 out of the q covariates are pertinent. Again, $x_i \sim \mathcal{N}_q(0, \frac{1}{4}I_q)$, where the factor $\frac{1}{4}$ is present so that the generated linear predictors $x_i^t\beta$ are distributed according to a standard normal distribution. For each simulation run, we minimize the information criterion under investigation, and force the intercept term to be in every model. Within each simulation run, AIC and BIC select one single best model, while for each one of the n_{test} observations in the test sample, possibly different models according to FIC_{MSE}, FIC_{MAE} and FIC_{ER} are selected. The forward search method as described in Section 4.2 has been used, and in each of those selected models we use the estimator $\hat{\mu}_{0,i} = \hat{\theta} + x_{0,i}^t \hat{\gamma}$. Its sign determines the predicted value of the corresponding binary $y_{0,i}$ values. We did experiments with $n_{\text{train}} = 50$ and 200, and $q = 5$ and 9.

For each separate observation $x_{0,i}$ in the test sample, with $1 \leq i \leq n_{\text{test}}$, we measure the performance of the model selection criteria via (a) the mean squared error of $\hat{\mu}_{0,i}$, (b) its mean average deviation, and (c) the error rate. The MSE 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 value for validation observation $x_{0,i}$ in simulation run j, and $\mu_{0,i,\text{true}}$

the corresponding true value. Similarly, the MAE is computed as

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

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

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

where $I(\cdot)$ is the indicator function. If the estimated and the true linear predictor have the same sign, they give a zero contribution to the sum in the above ER_i. Otherwise, they contribute to the error rate.

4.2 Further 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. Alternatively, we can apply a backward elimination procedure, starting with the full model, and eliminating in each step the variable which gives the largest reduction (or smallest increase) to the value of the information criterion. This will also lead to $q+1$ nested models as described above, from which we choose the model with the lowest value of 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 estimators in the different models. For each of the nested models obtained during the forward variable selection procedure, we compute the 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, BIC, 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 $\mu_0 = x_0^t\beta$ 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 might have 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 $\text{AIC}_S = -2\log L(\hat{\beta}_S) + 2(p + |S|)$, and $\text{BIC}_S = -2\log L(\hat{\beta}_S) + \log(n_{\text{train}})(p + |S|)$, with $L(\hat{\beta}_S)$ the likelihood of the estimated model indexed by S, and $|S|$ the number of elements in the subset S, such that lower values indicate better models.

4.3 Simulation results

This simulation results in $n_{\text{test}} = 500$ distinct values of the MSE, MAE and Error Rate, one for each observation in the test sample, for prediction based on a submodel selected by AIC, BIC, FIC_{MSE}, FIC_{MAE}, and FIC_{ER}. These values are also computed for the model-averaged predictions, discussed in Section 4.2. For the case $n_{\text{train}} = 50$ and $q = 5$, Table 1 presents the averages, after applying the log-transform to MSE and MAE, of the performance measures

over the $n_{\text{test}} = 500$ values, together with their standard error (SE). The log-transformation is applied to the MSE and the MAE, to make their distributions more symmetric. The boxplots in Figures 1 and 2 provide a graphical representation of these 500 values.

Please insert Figures 1 and 2, and Table 1 here.

First of all, we see from Table 1 that model averaging significantly improves the performance for the MSE and MAE. In terms of Error Rate, model averaging does not seem to give much improvement, but neither a worsening of the results obtained with single model selection. We see that FIC_{ER} gives the best results for the Error Rate, FIC_{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 FIC_{ER} is the error rate (to be more precise, the error rate of the limiting experiment), and FIC_{ER} selects the model having the smallest value of an approximation of this risk measure. It can be verified that the average Error Rate of the FIC_{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. The average error rates are computed over n_{test} outcomes, and differences among them have been tested for by performing multiple paired comparisons tests with Tukey's Honest Significant Difference method (e.g., Neter et al., 1996, page 725-732) and resulted in P-values < 0.01. Also, comparing with the results from the full model, given in the bottom line in Table 1, we see that FIC_{MSE} outperforms the full model in terms of MSE and MAE, and that FIC_{ER} does as good as the full model in terms of Error Rate. The models selected by FIC_{ER} however, generally have a small number of selected variables, and hence are much easier to interpret than the model which includes all variables.

The plots in Figure 1 show that FIC_{MSE} and FIC_{MAE} outperform the selection procedure based on AIC and BIC when using MSE and MAE as performance criterion. Again, one can show that these differences in average performance are also highly significant, and become

after model-averaging even more pronounced. This is as one should expect, since variable selection using FIC_{MSE} and FIC_{MAE} is aimed at choosing the "best" model as measured by the risks MSE and MAE. While FIC_{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.

Comparing FIC_{MSE} and FIC_{MAE} is more difficult. When selecting a single model, the MAE for estimates based on FIC_{MAE} is on average slightly worse than for FIC_{MSE}, although the difference is only minor. Note that at the finite-sample level there is no guarantee that the model selected using the FIC_{MAE} indeed yields the smallest Mean Absolute Errors. Moreover, the FIC is only estimating the limiting risk measures, and uncertainty from estimating population quantities needs to be taken into account. Most important, however, is that in this simulation setting, both FIC_{MSE} and FIC_{MAE} do better than AIC and BIC, both for model selection and model averaging.

Our simulations also indicated that increasing the number of variables q to 9, or increasing the training sample size to 200 does not change the above conclusions. Of course, for $n_{\text{train}} = 200$ all MSE/MAE will be lower than for a training sample size of 50. In Figure 2, boxplot representations of the n_{test} simulated error rates are given for the cases (i) $n_{\text{train}} = 50$ and $q = 5$ (ii) $n_{\text{train}} = 50$ and $q = 9$ and (iii) $n_{\text{train}} = 200$ and $q = 5$. Again we observe that FIC_{ER} performs the best on this criterion, especially for small training sample sizes ($n_{\text{train}} = 50$), and this remains true if we apply model averaging. We also observe that for the larger training sample sizes ($n_{\text{train}} = 200$), the performances of the different model selection methods are closer together. This is again as expected, since if n_{train} gets larger, the variance of the parameter estimators decreases.

5 Analysis of WESDR Data

In this section we perform model selection for the 1998 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 11 continuous covariates are 'rere' and 'lere', the refractive error in diopters for respectively the right and the left eye; 'reip' and 'leip', the internal eye pressure in mmHg for respectively 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 systolic and diastolic blood pressure in mmHg; 'bmi', the Body Mass Index, and 'pulse', the pulse rate in beats per 30 seconds. The 4 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.

When we fit a model including all the variables, we find that the following are significant at the 5% level: 'ddia', 'gly', 'urb' (in decreasing order of significance). Some pairs of variables are strongly correlated, for example 'lere' and 'rere' ($r = 0.869$), and 'reip' and 'leip' ($r = 0.872$). These four variables are also the ones with the largest Variance Inflation Factor (above the critical value 3), as computed by the `R` software package, following Davis, Hyde, Bangdiwala, and Nelson (1986), pp. 140–147. We refer to Klein et al. (1984) for further discussion of this data set. Given the high number of variables and the correlations among them, we want to select a subset of variables, most pertinent for predicting the response variable for a new patient.

We examine the predictive power of the models selected by the different selection criteria AIC, BIC, FIC_{MSE}, FIC_{MAE}, FIC_{ER}, as well as the model-averaged version by assessing their

error rates. Since the total number of all possible submodels amounts to 2^{15}, we carried out the model selection using a forward search procedure, as discussed in Section 4.2, to speed up the computation time. Also note that, since we work with real data for which the true value of the linear predictors 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. 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 and BIC counterparts. The McNemar test (e.g. Kuncheva 2004, page 13-15) reveals that in particular the difference with the AIC-selected model 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. If the "right" model would be within the class of allowed models, then this is presumably the best model to use for prediction. However, we do not believe that the "right" model does exist, only that some models are better than others, depending on the purpose of the analysis.

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 FIC_{MSE} and FIC_{ER}.

Please insert Table 3 here.

The 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 0.189. In fact, the model selected using FIC_{ER} ends up with this simple model in 46.3% of the cases. But, as follows from Table 2, the 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 add 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. Fitting a logistic regression model containing only the intercept, 'ddia' and 'gly', we find a cross-validated error rate of 0.184, still above the error rates found with the focussed information criteria. (Note that adding the third most

significant variable, 'urb', does not further improve the error rate). In Table 3, the variables being selected first in the forward procedure by AIC and BIC are also reported. We see that BIC only selects 'ddia' and 'gly', while the model finally selected by the AIC criterion contains 7 variables.

Variable selection based on FIC_{ER} includes the variable 'gly' much more often for groups B and D than for groups A and C (see Table 3). Hence, there is some indication that the glycosylated hemoglobin level is, from a predictive point of view, less important 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). If a full model approach is opted for, it might be advisable to include an interaction term between the two variables 'gly' and 'ins'.

6 Concluding comments

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 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 better with respect to their specific risk measure (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, a gain in predictive performance is expected.

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: We wish to thank Dr. Ronald Klein for kindly giving permission to use the WESDR data. Secondly, we thank Prof. Nils Lid Hjort for his insightful comments and suggestions. We also thank the reviewers for their constructive comments and suggestions for improvement. This research 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.

Davis C. E., Hyde J. E., Bangdiwala S. I. and Nelson J. J. (1986). An example of dependencies among variables in a conditional logistic regression. *Modern Statistical Methods in Chronic Disease Epidemiology*, Eds. S.H. Moolgavkar and R.L. Prentice, New York:

Wiley.

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

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

Hastie, T., Tibshirani, R. and 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 retinopathy: 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.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Statistical Models: Fourth Edition.* Irwin.

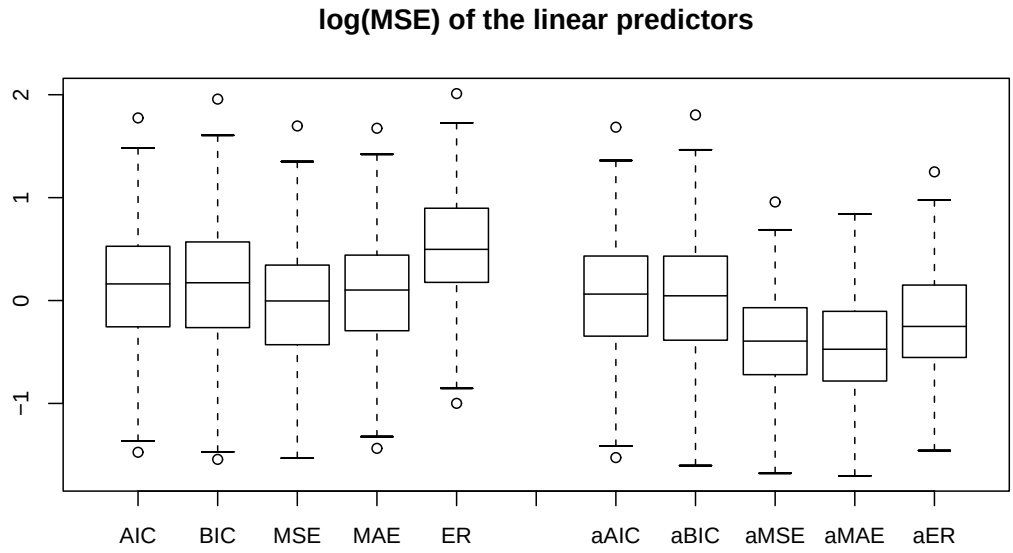
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

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.

Figure captions

Figure 1. Boxplots of the log(MSE) and log(MAE) of the 500 observations to predict in the test sample for the sampling scheme with $n_{\text{train}} = 50$ and $q = 5$. The MSE and MAE have been simulated for estimators of a model selected by the criteria AIC, BIC, FIC_{MSE}, FIC_{MAE}, or FIC_{ER}, as well as for the model averaged versions of the estimators (indicated by the prefix "a").

Figure 2. Boxplots of the Error Rates of the 500 observations to predict in the test sample. These Error Rates have been simulated for estimators of a model selected by the criteria AIC, BIC, FIC_{MSE}, FIC_{MAE}, or FIC_{ER}, as well as for the model averaged versions of the estimators (indicated by the prefix "a"). In the top panel (a) $n_{\text{train}} = 50$, and $q = 5$ variables, in (b) $n_{\text{train}} = 50$, and $q = 9$ variables, and in panel (c) $n_{\text{train}} = 200$, and $q = 5$ variables.


log(MSE) of the linear predictors
2
1
0
−1
AIC
BIC
MSE
MAE
ER
aAIC
aBIC
aMSE
aMAE
aER

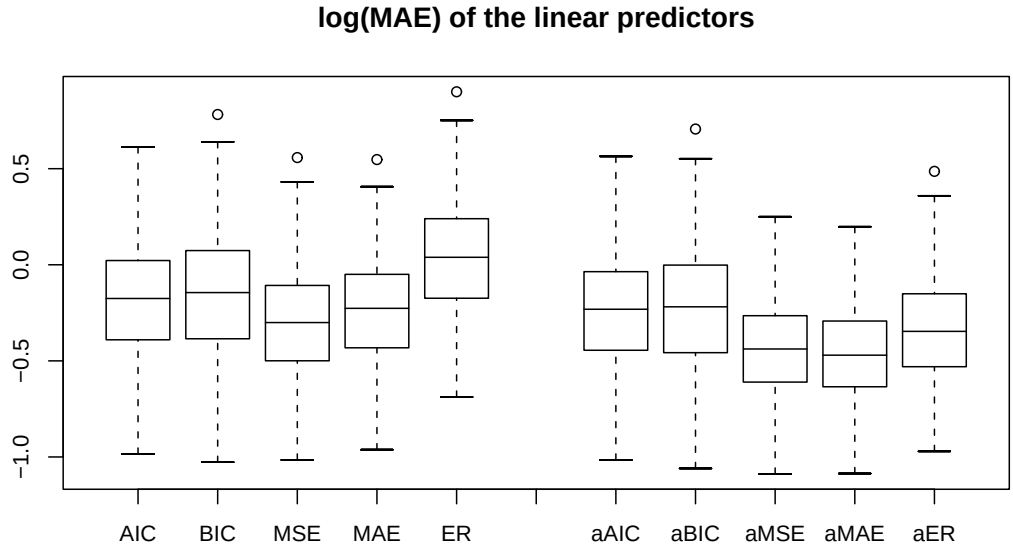

log(MAE) of the linear predictors
0.5
0.0
−0.5
−1.0
AIC
BIC
MSE
MAE
ER
aAIC
aBIC
aMSE
aMAE
aER

Figure 1:

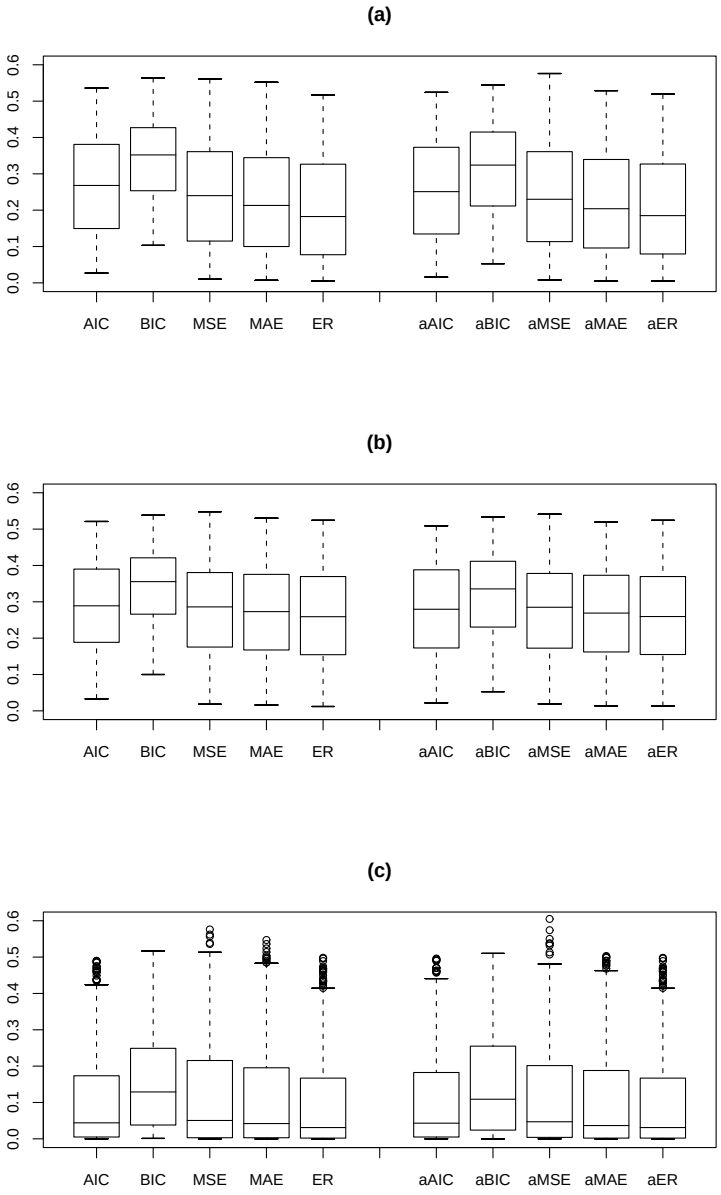

(a)
0.0 0.1 0.2 0.3 0.4 0.5 0.6
AIC BIC MSE MAE ER aAIC aBIC aMSE aMAE aER
(b)
0.0 0.1 0.2 0.3 0.4 0.5 0.6
AIC BIC MSE MAE ER aAIC aBIC aMSE aMAE aER
(c)
0.0 0.1 0.2 0.3 0.4 0.5 0.6
AIC BIC MSE MAE ER aAIC aBIC aMSE aMAE aER

Figure 2:

Table 1: Average values, together with their standard errors (SE), of the log(MSE), log(MAE) and Error Rates over the 500 observations to predict in the test sample for the sampling scheme with $n_{\text{train}} = 50$ and $q = 5$. The MSE, MAE, and Error rates have been simulated for estimators of a model selected by the criteria AIC, FIC_{MSE}, FIC_{MAE}, and FIC_{ER}, as well as for the model averaged versions of the estimators (indicated by the prefix "a").

	log(MSE)		log(MAE)		Error Rate ($\times 10^{-2}$)	
Criterion	Average	SE	Average	SE	Average	SE
AIC	0.141	0.025	-0.182	0.013	26.62	0.60
BIC	0.153	0.027	-0.152	0.014	33.89	0.48
FIC_{MSE}	-0.026	0.024	-0.298	0.013	24.65	0.64
FIC_{MAE}	0.085	0.024	-0.238	0.012	22.87	0.65
FIC_{ER}	0.507	0.024	0.034	0.013	20.75	0.65
aAIC	0.045	0.025	-0.238	0.013	25.45	0.62
aBIC	0.025	0.026	-0.226	0.014	31.14	0.55
$a\text{FIC}_{\text{MSE}}$	-0.402	0.021	-0.438	0.011	24.23	0.64
$a\text{FIC}_{\text{MAE}}$	-0.454	0.021	-0.467	0.011	22.34	0.64
$a\text{FIC}_{\text{ER}}$	-0.220	0.023	-0.341	0.013	20.91	0.64
full model	0.065	0.024	-0.253	0.012	20.75	0.65

Table 2: Error rates for the WESDR data, obtained via cross-validation. The models are selected using AIC, BIC FIC_{MSE}, FIC_{MAE} FIC_{ER} and also results for the model-averaged estimates are reported.

Method	AIC	BIC	FIC_{MSE}	FIC_{MAE}	FIC_{ER}
Error rate (no model averaging)	0.198	0.184	0.174	0.174	0.177
Error rate (after model averaging)	0.194	0.188	0.171	0.174	0.174

Table 3: Model selection methods FIC_{MSE} and FIC_{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. For completeness, the last 2 rows show the first four variables considered for inclusion by AIC and BIC, and whether they have been selected ("yes") or not ("no").

	Group	Variable 1		Variable 2		Variable 3		Variable 4	
FIC_{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%
FIC_{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%
AIC		ddia	yes	gly	yes	bmi	yes	pulse	yes
BIC		ddia	yes	gly	yes	bmi	no	pulse	no