Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room

Green, Michael; Björk, Jonas; Forberg, Jakob; Ekelund, Ulf; Edenbrandt, Lars; Ohlsson, Mattias

Published in:
Artificial Intelligence in Medicine

DOI:
10.1016/j.artmed.2006.07.006

2006

Artificial Intelligence in Medicine (2006) **38**, 305—318

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room

Michael Green [a,*], Jonas Björk [b], Jakob Forberg [c], Ulf Ekelund [c], Lars Edenbrandt [d], Mattias Ohlsson [a]

[a] *Department of Theoretical Physics, Lund University, Sölvegatan 14A, SE-22362 Lund, Sweden*
[b] *Competence Centre for Clinical Research, Lund University Hospital, SE-22185 Lund, Sweden*
[c] *Department of Emergency Medicine, Lund University Hospital, SE-22185 Lund, Sweden*
[d] *Department of Clinical Physiology, Malmö University Hospital, SE-20502 Malmö, Sweden*

Received 21 November 2005; received in revised form 5 July 2006; accepted 12 July 2006

KEYWORDS
Artificial neural networks;
Ensemble methods;
Logistic regression;
Acute coronary syndrome;
Acute myocardial infarction;
Clinical decision support

Summary

Objective: Patients with suspicion of acute coronary syndrome (ACS) are difficult to diagnose and they represent a very heterogeneous group. Some require immediate treatment while others, with only minor disorders, may be sent home. Detecting ACS patients using a machine learning approach would be advantageous in many situations.
Methods and materials: Artificial neural network (ANN) ensembles and logistic regression models were trained on data from 634 patients presenting an emergency department with chest pain. Only data immediately available at patient presentation were used, including electrocardiogram (ECG) data. The models were analyzed using receiver operating characteristics (ROC) curve analysis, calibration assessments, inter- and intra-method variations. Effective odds ratios for the ANN ensembles were compared with the odds ratios obtained from the logistic model.
Results: The ANN ensemble approach together with ECG data preprocessed using principal component analysis resulted in an area under the ROC curve of 80%. At the sensitivity of 95% the specificity was 41%, corresponding to a negative predictive value of 97%, given the ACS prevalence of 21%. Adding clinical data available at presentation did not improve the ANN ensemble performance. Using the area under the ROC curve and model calibration as measures of performance we found an advantage using the ANN ensemble models compared to the logistic regression models.

* Corresponding author. Tel.: +46 222 34 94; fax: +46 222 96 86.
E-mail address: michael@thep.lu.se (M. Green).

0933-3657/$ — see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2006.07.006

Conclusion: Clinically, a prediction model of the present type, combined with the judgment of trained emergency department personnel, could be useful for the early discharge of chest pain patients in populations with a low prevalence of ACS.

1. Introduction

Patients who present at the emergency department (ED) with chest pain or other symptoms suspicious of myocardial infarction (AMI) or unstable angina pectoris (i.e. acute coronary syndrome, ACS) are common and represent a heterogeneous group. Some have an AMI with a high risk of life-threatening complications, whereas others have completely benign disorders which may safely be evaluated on an out-patient basis. Since our ability to diagnose ACS in the ED remains poor, and since the consequences of a missed ACS can be disastrous, there is a large overadmission to in-hospital care; some 7 out of 10 patients admitted with a suspicion of ACS prove not have it [1,2].

A number of methods have been developed to support the physicians in their decision making regarding patients presenting to the ED with chest pain [3—9]. Goldman et al. [3] developed a statistical model to estimate the relative risk of major events within 72 h after arrival at the ED. The independent variables used included age, gender and electrocardiographic (ECG) findings, all available at presentation. Another model, the ACI-TIPI [4] was developed to assist triage decisions regarding patients with symptoms of acute cardiac ischemia. This model, using only a few factors (both clinical and ECG), was able to significantly reduce hospitalizations for ED patients without acute cardiac ischemia. In a recent study by Harrison et al. [7] approximately 3000 ACS patients from three different hospitals were analyzed with very good results, using as few as eight features. They obtained an area under the receiver operating characteristics (ROC) curve as high as 98%. An example of ACS prediction can also be found in the work of Xue et al. [6] where a hybrid machine learning approach was used, combining artificial neural networks (ANN) and decision trees. There are also a number of approaches that have been developed to predict the presence of AMI based on a full range of clinical data [10—13] and data limited to the 12-lead ECG only [14,15]. Many of these methods used ANN as the classification tool. The performance is usually good compared to interpretation made by experienced physicians.

ANN represents a machine learning tool that has turned out to be useful for complex pattern recognition problems. ANN is also widely used for medical applications (see e.g. [16]). Ensemble learning for ANN is standard procedure to increase the generalization performance by combining several individual networks trained on the same task. The ensemble approach has been justified both theoretically [17,18] and empirically [19]. Combining the outputs is clearly only relevant when they disagree on some or several of the samples. The most simple method for creating diverse ensemble members is to train each network using randomly initialized weights (also known as injecting randomness). A more elaborate approach is to train the different networks on different subsets of the training set. An example is bagging [20] where each training set is created by resampling (with replacement) the original one, with uniform probability. Cross-splitting [18] is another ensemble creation technique that has performed well in connection with ACS prediction [8].

Comparing ANN models with standard statistical generalized linear models such as logistic regression is an important step in the development procedure. If the results show that the gain of using a non-linear model, such as the ANN, is limited, one should usually go for the less complicated model. Logistic regression always has the nice property of being fully interpretable which can be used to provide feed-back to the user. When performing this comparison it is always important to use more than one measure of performance, since there are several aspects of what is good performance [21].

The aims in this study were two-fold. The first aim was to construct an ACS prediction model for our study population and explore to what extent we can confirm previous results obtained for other ACS study populations. Part of this aim was also to identify relevant clinical input factors for the ACS prediction models using an effective odds ratio approach. The second aim was to conduct a detailed comparison between ANN and logistic regression models. In this comparison, we used two common techniques for ANN ensemble training together with a single ANN approach. The measures of performance were area under the ROC curve, χ^2 calibration statistics and Pearson correlations for intra- and inter-method variations.

Table 1 Characteristics of the independent variables used to train the ACS prediction models

Input variable	No miss., *n*	ACS, *n* (%)	No ACS, *n* (%)
Age	—	70.1[a](13.2)[b]	61.3[a](18.0)[b c]
Gender	—		
Male		83 (63.8)	279 (55.4)[c]
Female		47 (36.2)	225 (44.6)
Diastolic blood pressure	15	83.9[a](14.9)[b]	82.7[a](12.4)[b c]
Systolic blood pressure	8	148.5[a](29.6)[b]	142.2[a](24.0)[b]
Heart rate	2	79.4[a](22.0)[b]	78.1[a](18.1)[b]
Smoking status	—		
Current		29 (22.3)	98 (19.4)
Not current/unknown		101 (77.7)	406 (80.6)
Hypertension	—		
Yes		47 (36.2)	114 (22.6)[c]
No/unknown		83 (63.8)	390 (77.4)
Diabetes	—		
Yes		19 (14.6)	57 (11.3)
No		111 (85.4)	447 (88.7)
Medication	—		
Yes		82 (63.1)	263 (52.2)
No		48 (36.9)	241 (47.8)
Angina pectoris	2		
Yes, ≤ 1 month		4 (3.1)	5 (1.0)[c]
Yes, > 1 month		56 (43.8)	174 (34.5)
No		68 (53.1)	325 (64.5)
Congestive heart failure	—		
Yes		20 (15.4)	79 (15.7)[c]
No		110 (84.6)	425 (84.3)
Chest discomfort at presentation	—		
Yes		85 (65.4)	238 (47.2)[c]
No		45 (34.6)	266 (52.8)
Symptom duration	2		
0—6 h		100 (76.9)	263 (52.2)[c]
7—12 h		16 (12.3)	59 (11.7)[c]
13—24 h		4 (3.1)	42 (8.3)
>24 h		10 (7.7)	140 (27.8)
Tachypnea	—		
Yes		13 (10.0)	27 (5.4)
No		117 (90.0)	477 (94.6)
Lung rales	—		
Yes		12 (9.2)	23 (4.6)
No		118 (90.8)	481 (95.4)
Previous myocardial infarction	—		
Yes, ≤ 6 months		13 (10.0)	19 (3.8)[c]
Yes, > 6 months		37 (28.5)	107 (21.2)[c]
No		80 (61.5)	378 (75.0)
Previous PTCA	—		
Yes		4 (3.1)	21 (4.2)[c]
No		126 (96.9)	483 (95.8)
Previous CABG	—		
Yes		10 (7.7)	55 (10.9)[c]
No		120 (92.3)	449 (89.1)

There are 130 cases of ACS and 504 cases without ACS. The second column shows the number of missing values for each variable, where '—' indicates no missing value. The last two columns shows the number of patients (percentage) in each category. For continuous variables the mean (S.D.) is presented. Also, footnote 'c' is used to indicate if a variable is part of the simplified logistic regression model.

[a] Mean.

[b] S.D.

[c] Clinical variables used in the simplified logistic regression model.

2. Materials and methods

2.1. Study population

This study is based on patients with chest pain attending the ED of Lund University Hospital, Sweden, from 1st July to 20th November 1997. Six hundred sixty-five consecutive visits for which electronic ECG data could be retrieved were included. To have as independent data as possible, some visits were removed such that a criterion of atleast 20 days between two consecutive visits, for a given patient, was fulfilled. This reduced the dataset to 634 visits, where 130 patients were diagnosed with ACS and 504 with no ACS. ECG data comprised the 12-lead ECG, recorded using computerized electrocardiographs (Siemens-Elema AB, Solna, Sweden). Table 1 shows the clinical variables used in this study. Missing values were substituted by the most common category for categorical variables and the mean value for continuous variables.

ECG data were reduced to smaller sets of more effective variables before entered into the classification models. The reduction was accomplished using principal component analysis (PCA). Prior to this analysis the measurements were grouped into the following six sets of measurements namely: QRS area (total area of the QRS complex), QRS duration, QRS amplitudes, ST amplitudes (ST-amp, ST-amp 2/8 and ST-amp 3/8), ST slope (the slope at the beginning of the ST segment) and positive/negative T amplitudes. The ST amplitudes 2/8 and 3/8 were obtained by dividing the interval between ST-J point and the end of the T wave into eight parts of equal duration. The amplitudes at the end of the second and third interval were denoted ST amplitude 2/8 and 3/8, respectively. Each of these six sets were then subject to a principal component analysis reduction, e.g. the 12 ST slope variables (one from each lead) were reduced to two variables. The final ECG data set, to be used for the ANN training, consisted of a selection [22] of 16 PCA variables.

The diagnosis of ACS is defined as one of the following discharge diagnoses for the patient: AMI and unstable angina pectoris. The discharge diagnoses were made by the attending senior ward physicians and also reviewed by an experienced research nurse. AMI was defined by the WHO criteria [23] where the biochemical criterion was atleast one measurement of CK-MB > 10 μg/l or Troponin T > 0.1 μg/l. The criteria for unstable angina were (i) observed with (ii) and/or (iii):

(i) Ischemic symptoms: chest pain $>$ 15 min, syncope, acute heart failure or pulmonary oedema.

(ii) Electrocardiogram (ECG) changes: transient or persisting ST segment depression ($\geq$ 1 mm) and/or T-wave inversion ($\geq$ 1 mm) without developing Q waves or loss of R wave height.

(iii) Biochemical markers: CK-MB 5–10 μg/l or Troponin T 0.05–0.1 μg/l.

Table 2 Characteristics of the ECGs recorded on the patients

ECG finding	ACS *n* (%)	No ACS *n* (%)
ST-elevation	52 (40.0)	80 (15.9)
ST-depression	52 (40.0)	59 (11.7)
T-wave inversion	74 (56.9)	189 (37.5)

There are 130 cases of ACS and 504 cases without ACS. ST-elevation was defined as ST amplitude $\geq$ 1 mm in two or more contiguous leads, whereas ST-depression was defined as a negative ST amplitude $\geq$ 1 mm in any lead. T-wave depression was defined as a negative T-wave ($\geq$ 1 mm) with a predominant R-wave.

The non-ACS cases consisted of patients with the diagnosis of stable and suspected angina pectoris, together with the category "other diagnosis". Out of the 504 non-ACS cases, 271 had discharge diagnoses other than stable or suspected angina pectoris. Table 2 shows common ECG characteristics for both the ACS cases and the non-ACS cases, obtained by the lead measurements.

2.2. Artificial neural networks

We considered ANN in the form of feed-forward multilayer perceptrons (MLP) with one hidden layer and no direct input–output connections. The hidden unit activation function was the hyperbolic tangents and the output activation function was the standard logistic function. We used the cross-entropy error function for two classes. In addition, we introduced a weight elimination term E_{reg} [24], controlled by a tunable parameter λ, to possibly regularize the network:

$$E_{\text{reg}} = \lambda \sum_i \frac{\beta_i^2}{1+\beta_i^2}$$

where the sum runs over all weights in the MLP, except threshold weights. The total error is the sum of the cross-entropy part and E_{reg} for the case when using regularized MLPs. The minimization of the error function was accomplished using the gradient descent method.

Among several existing methods for constructing ensembles, such as voting and boosting (see e.g. [25]) we have used two methods: the common bagging method [20] and *S*-fold cross-splitting [18,8]. In bagging one starts with a given training set and then creates new training sets by resampling, with replacement, the original one. Thus, the bagging ensemble contains MLPs trained on *bootstrap* samples of the original training set. The ensemble output t^{ens} is simply computed as the mean of the individual ensemble members, i.e.,

$$t^{\text{ens}} = \frac{1}{C}\sum_{n=C}^{C} t_n \quad (1)$$

where t_n is the output of the n: th MLP in the ensemble and C is the bagging ensemble size.

Another way to create diverse training sets is to randomly partition the dataset into S bins. One can then create S slightly different training sets by excluding one of the parts each time. This procedure can be repeated N times to create $N \times S$ different but similar training sets. By training an MLP on each of these training sets we can create a pool of MLPs that can be combined into a $N \times S$ cross-splitting ensemble. As for bagging the ensemble output is computed as the mean over the $N \times S$ MLP outputs (see Eq. (1)). Clearly, the difference between the training sets will increase if fewer bins are used, as a larger fraction of the original training set is excluded each time. For the efficiency of the ensemble we therefore used $S = 2$, supported by the findings in Green et al. [8]. This approach to ensemble creation can be found in the work of Krogh et al. [18], but used in a different context.

The ensemble size, C for bagging and $N \times S$ for cross-splitting, influences the performance of the ensemble method compared to single MLP classifiers. In this study we used an ensemble size of 25 for the model selection and 50 for the final test runs. Both sizes are reasonable according to numerical studies (see e.g. [19,26]).

2.3. Ensemble model selection

Even though the use of ensembles decreases the usual negative effect of overtraining, one must perform model selection for the ensemble. We use the standard K-fold cross-validation procedure to estimate the generalization performance. However, to actually validate the ensemble, each training group in the K-fold cross-validation procedure is used to train an ensemble with either bagging or S-fold cross-splitting. Fig. 1 summarizes the procedure used for performing ensemble model selection. Model selection is performed, using a grid search, over parameters λ and the number of hidden units in the ANN.

Alternative procedures can be used with the S-fold cross-splitting ensemble, which combines both the cross-validation and the ensembles creation [8]. However to accurately validate both the bagging and the S-fold cross-splitting ensemble we used the above procedure even though it is more costly in terms of CPU-time.

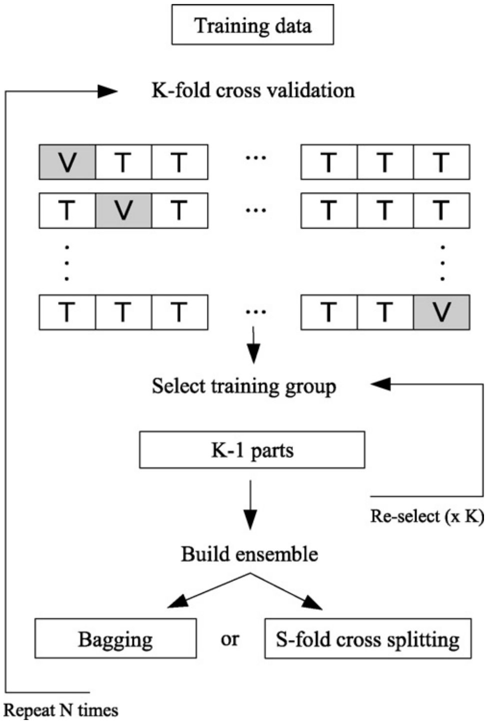

Figure 1 Ensemble model selection procedure. A given training data set was split into several training/validation parts using K-fold cross-validation. Each of these smaller training sets (T) were then used to create an ANN ensemble and the corresponding validation set (V) was used for validation. For each K-fold cross-validation split, K ensembles were created which resulted in K validation results. The whole procedure was repeated N times with different random K-fold cross-validation splits.

2.4. Multiple logistic regression

Multiple logistic regression [27] was also used to predict the probability of ACS. Both full logistic regression models, using the same inputs as the ANN models, and a simplified model using only clinical input data were trained. The clinical input variables used for the simplified logistic regression model can be found in Table 1.

The optimization procedure for the simplified logistic regression model was as follows: starting with the full multivariate model with all independent variables included, we excluded one insignificant independent variable at a time, starting with the variable with highest p-value, until only significant and important predictors remained. Categorical variables with more than two categories were kept in the model if the odds ratio associated with any of the categories was significant. The statistical power to detect associations between some of the rare but possibly important clinical characteristics was low. Thus, variables with estimated odds ratio of atleast 2.5 (or, equivalently, atmost 0.4) were considered as important predictors and kept in the model even if they were not statistically significant. In order to simplify the final model, categories with odds ratios close to one were collapsed with the reference category for that variable. Similarly, unknown response to one of the variables (hypertension) was also added to the reference category.

2.5. Statistical analysis

2.5.1. Effective odds ratios

To discern the information content in each of the ANN input features we considered *effective odds ratios*. Odds ratio is the ratio between the odds for an event when a feature is present and the odds for an event when that feature is absent. Odds ratios are well known in the statistical community but cannot be used in conjunction with ANN since the output of an ANN is a non-linear function of the inputs. Odds ratios are defined as:

$$\mathrm{OR} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{p_1(1-p_0)}{p_0(1-p_1)} \qquad (2)$$

where p_1 is the risk of an event for a patient with a certain feature and p_0 is the risk for the patient without that certain feature. In generalized linear models, such as the logistic regression model used in this study, the odds ratio for a particular feature is e^w, where w is the weight for this particular feature. In an ANN we have a non-linear function in the exponent which depends on all other input features in the ANN. However, it is possible to calculate an effective odds ratio by averaging expression (2) over all patients [28].

For the logistic regression model there is an alternative interpretation of the odds ratio for a specific feature. The logistic standard bare model can be described by the following relation:

$$y = \sum_{i=1}^{m} x_i \omega_i + \omega_0$$

where y is the log odds of an event, given the input $(x_1, x_2, \ldots, x_m)$. If we take the derivative of this relation with respect to a certain feature x_i we end up with:

$$\frac{\partial y}{\partial x_i} = \omega_i = \log(\mathrm{OR}_{x_i}) \qquad (3)$$

In other words, we can interpret the derivative with respect to a feature x_i as the log odds ratio for that feature. We can easily generalize this measure to the ANN case. However, the resulting expression will depend on the other input features via the hidden layer function. We can consider odds ratios for an ANN as either the effective odds ratio where we average expression (2) over all patients, or we can use the derivative interpretation, by averaging expression (3). It is not obvious which one provides the best approximation of odds ratios for the ANN. In this study we used the former approach.

2.5.2. Model calibration

Model calibration, which is a comparison between the observed and predicted ACS risk, was evaluated using the Hosmer—Lemeshow goodness-of-fit test [29], which is given by,

$$\chi^2 = \sum_{j=1}^{G} \frac{(o_j - n_j \bar{\pi}_j)^2}{n_j \bar{\pi}_j (1 - \bar{\pi}_j)}$$

In this expression o_j is the number of observed ACS cases in bin j, $\bar{\pi}_j$ the mean average predicted ACS risk in bin j, and G the number of bins meanwhile n_j is the number of samples in the bin. This test follows the χ^2 statistics with $(G-2)$ degrees of freedom. In this study we have used 10 bins of equal size. The resulting χ^2 statistic is used to indicate non-significant differences ($p > 0.05$) between observed and predicted ACS.

2.6. Performance estimation

In addition to the calibration assessment we also constructed ROC curves for all methods. The area under the ROC curve provides yet another (popular) measure of performance. It has the interpretation of the probability that a randomly chosen patient with ACS has a larger predicted ACS risk than a randomly chosen patient without ACS (see e.g. [30]). From the ROC curve we also accessed the specificity at a level of 95% sensitivity. This somewhat arbitrary level was chosen because with current standard evaluation, some 2—5% of the ACS patients are erroneously discharged from the ED, which implies a sensitivity of atleast 95% for the routine ED work-up.

To estimate the generalization performance of the tested models we used a five-fold cross-testing procedure, repeated 20 times, resulting in 100 test sets on which the area under the ROC curve was calculated. The procedure is similar to the cross-validation method used for model selection and is accomplished by dividing the data set into five parts of (approximately) equal size. An ACS prediction model is constructed on all parts except one, which is used as the independent test set. The median of the 100 ROC areas is used as the test performance for a given model and selection of independent variables.

An alternative approach to measure the generalization performance is to make an ensemble of the test ACS predictions. This is accomplished by computing the average ACS probability for each patient taken over the 20 cross-splittings defined above. The end result is a single list of test ACS probabilities, comprising the full data set, and its corresponding ROC curve. The 100 test set predictions, for a given particular model, is thus transformed into one set of test predictions, defined as the *full test ensemble*. One would expect this approach to produce an estimation of the generalization performance that is above the one given by the median of the 100 single test results since there is yet another ensemble effect to account for. Furthermore, using the full test ensemble enables a straightforward statistical comparison between different ROC curves and their areas. Associated *p*-values for ROC area differences using the full test ensemble were calculated using a permutation test (see e.g. [31]).

2.7. Software

In this study we used the SAS system to build and develop the logistic regression models meanwhile a C++ based software package was used to build the ANN models. The statistical comparisons were conducted using custom made Perl scripts.

3. Results

The test ROC areas obtained for the different methods and different combinations of independent variables are summarized in Table 3. For each method the ROC area is given both as the median area of the 100 test sets and as the single area of the full test set ensemble.

Table 3 Test ROC areas obtained from the different methods

Model	Number of variables (categories[a] + continuous)	Test ROC area (%)
ANN bagging ensemble		
Clinical + ECG data	38	79.1 (69.2, 86.2)
		80.1 (76.2, 84.2)
ECG data	16	79.8 (69.2, 88.5)
		81.1 (77.1, 85.2)
Clinical data	22	75.3 (67.2, 83.0)
		76.0 (71.8, 80.4)
ANN cross-splitting ensemble		
Clinical + ECG data	38	78.7 (68.6, 86.5)
		80.0 (76.1, 84.0)
ECG data	16	80.2 (70.7, 89.2)
		81.0 (77.1, 85.2)
Clinical data	22	75.1 (67.0, 82.6)
		75.3 (70.9, 79.8)
ANN single MLP		
Clinical + ECG data	38	76.3 (65.3, 83.7)
		77.1 (72.7, 81.6)
ECG data	16	76.0 (60.0, 87.1)
		80.0 (76.0, 84.2)
Clinical data	22	72.6 (64.9, 80.7)
		73.3 (68.6, 78.1)
Multiple logistic regression (no interaction)		
Clinical + ECG data	38	75.7 (63.5, 84.2)
		76.4 (71.8, 80.9)
ECG data	16	70.5 (54.2, 81.2)
		71.0 (65.8, 76.2)
Clinical data	22	72.5 (64.6, 81.7)
		73.1 (68.4, 78.0)
Multiple logistic regression (simplified)		
Clinical data	13	75.2 (66.4, 82.8)
		75.1 (70.7, 79.7)

For each method two estimations of the generalization performance are presented. The first line corresponds to the median (2.5, 97.5 percentiles) over the 100 test sets defined by the cross-testing procedure. The second line is the ROC area (95% confidence bounds) from the full test set ensemble.

[a] The base categories are not counted.

The best areas were obtained using the ANN ensemble approach with ECG data, 79.8% and 80.2% (median values) for the bagging and the cross-splitting ensemble, respectively. Adding clinical data to the ANN models did not improve the performance, there was actually a slight decrease of the performance (79.1% and 78.7%), although not significant. Comparing the two ANN ensemble creation methods, it is apparent that both methods yielded similar results. The logistic regression model using both ECG and clinical data received an area of 75.7%. Using only ECG data in the logistic model the results dropped to only 70.5%, indicating the presence non-linearities in the ECG data that the logistic regression model could not capture. Comparing the logistic regression models, built on clinical data alone, the simplified model, using feature selection, and the normal model, with all features present, received an ROC area of 75.2% and 72.5%, respectively.

Using the full test ensemble when measuring the performance allows for a (statistical) comparison of two ROC curves. As can be seen in Table 3 there was an overall increase of the performance using the full test ensemble (except for the simplified logistic model) and this is most certainly due to the ensemble-averaging effect. The difference was significant ($p = 0.05$) when comparing the ANN bagging ensemble trained with clinical data only (76.0%) and ECG data only (81.1%). For the cross-splitting ensemble the corresponding (significant different) areas were 75.3% and 81.0% ($p = 0.03$). Using the simplified logistic regression model, where each non-significant input feature was removed, resulted in an ROC area of 75.1%. The logistic regression model with all features present performed worse, receiving an ROC area of 73.1% ($p = 0.02$). Also including ECG data in the logistic regression model did not significantly improve the performance compared to the simplified model based on clinical data only. It is also interesting to compare sensitivity and specificity values for the different methods. Fig. 2 shows the ROC curve for the full test ensemble using the ANN bagging ensemble and the logistic regression method. At the sensitivity level of 95% we obtained a specificity of 41.1% and 33.7% for the ANN and the logistic model, respec-

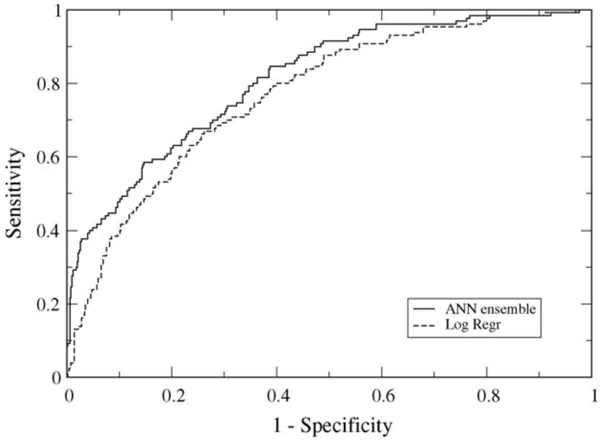

Figure 2 The ROC curves for the best ANN ensemble and the best logistic regression model using the full test ensemble. The areas under the curves were 81.1% and 76.4%, respectively. The difference was significant ($p = 0.03$).

Table 4 Test χ^2 calibration and intra-Pearson correlation values obtained from the different methods

Model	Calibration (χ^2)	Pearson correlation
ANN bagging ensemble		
Clinical + ECG data	14.5 (3.5, 58.8)	0.88 (0.85, 0.90)
ECG data	12.5 (3.2, 47.6)	0.85 (0.81, 0.88)
Clinical data	11.7 (4.1, 35.3)	0.92 (0.90, 0.93)
ANN cross-splitting ensemble		
Clinical + ECG data	13.6 (4.4, 65.3)	0.89 (0.86, 0.91)
ECG data	11.8 (3.6, 24.9)	0.85 (0.82, 0.88)
Clinical data	11.6 (3.2, 40.8)	0.93 (0.91, 0.94)
ANN single MLP		
Clinical + ECG data	15.7 (4.2, 65.2)	0.88 (0.85, 0.91)
ECG data	40.2 (7.3, 436.5)	0.69 (0.59, 0.78)
Clinical data	11.5 (3.5, 44.1)	0.93 (0.87, 0.95)
Multiple logistic regression		
Clinical + ECG data	24.8 (6.9, 93.6)	0.88 (0.84, 0.90)
ECG data	17.1 (3.9, 67.2)	0.85 (0.80, 0.89)
Clinical data	12.8 (4.5, 45.3)	0.93 (0.91, 0.95)
Multiple logistic regression (simplified)		
Clinical data	11.7 (3.6, 39.6)	0.96 (0.94, 0.97)

The values are presented as median (2.5, 97.5 percentiles) over the 100 test sets defined by the cross-testing procedure for the calibration assessment. Pearson correlation values are median (2.5, 97.5 percentiles) over all full test split pairs.

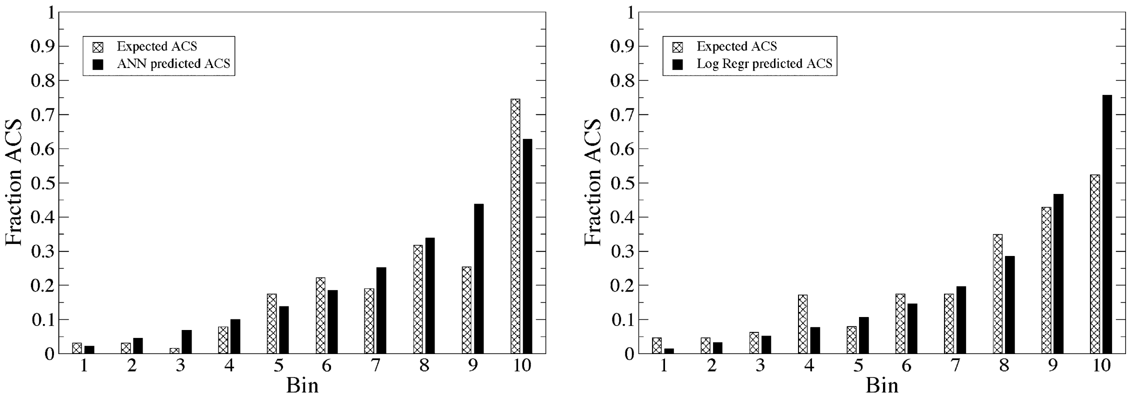

Figure 3 This figure shows the expected and the predicted fraction of ACS for patients in the full test ensemble. Left and right figure are the ANN ensemble, trained on ECG data only, and the logistic regression model, trained on both ECG and clinical data, respectively.

tively. With the prevalence of 20.5% ACS in this study population this corresponds to a negative predictive value of 97.2% (96.1%) and a positive predictive value of 29.5% (25.8%) for the ANN ensemble (logistic regression) method.

3.1. Calibration comparison

The degree of calibration for the different methods was quantified using the Hosmer—Lemeshow goodness-of-fit test [29]. The results are presented in Table 4. Comparing the best models (cross-splitting ensemble and logistic regression) we obtained χ^2 values of 11.8 and 24.8, respectively. Both values, taken as the median over the 100 test sets, corresponds to p-values of 0.16 and 0.002. We thus conclude that the best logistic regression model was not calibrated, meanwhile the ANN model was. Moreover, we see that the most calibrated model was the single MLP with a χ^2 and a p-value of 11.5 and 0.17, respectively. Generally models trained with only clinical data received the best calibration scores. The overall worse calibrated model was the single MLP model trained using only ECG data ($\chi^2 = 40.2$). An illustration of the degree of calibration in the full test ensemble is presented in Fig. 3 where the solid bars represent the predicted fraction of ACS meanwhile the textured bars represents the true fraction of ACS.

3.2. Scatter plots

Although the ROC area and the calibration comparison may reveal differences between the logistic regression and the ANN ensemble model, they are not useful for detecting differences on a patient per patient basis. It is therefore interesting to look at ordinary scatter plots, both for intra- and inter-method comparisons. To quantify the degree of correlation in the scatter plots we used the Pearson correlation coefficient. Results for the intra-method correlations can be found in Table 4. The simplified logistic regression model obtained the largest correlation coefficient (0.96). Generally methods trained with only clinical data had smaller intra-variations compared to method trained with ECG information. Comparing the best ANN and logistic regression model according to Table 3 we can conclude that the ANN had larger intra-method variations (0.85 compared to 0.88 for the logistic regression model). Fig. 4 shows the scatter plots for these two models, where the test splits are chosen as to correspond to median Pearson correlation values. Thus, the scatter plots in Fig. 4 represents typical intra-variations in the 20× five-fold cross-testing scheme for the two models.

For inter-method comparisons, we first looked at the best ANN model and the best logistic regression model according to the ROC area (see Table 3). The median Pearson correlation coefficient for all inter-method test split pairs was 0.59 and Fig. 5 (left part) shows a corresponding scatter plot. Since there was an ROC area difference of 4.5% between the two models (80.2% compared to 75.7%) one would expect some inter-method differences, but the scatter plot shows a large variation for many patients.

It is also interesting to compare ANN and logistic regression models that had almost the same ROC area and calibration statistics. The bagging ensemble trained on clinical data obtained an ROC area of 75.3% and calibration χ^2 of 11.7. The corresponding numbers for the simplified logistic regression model was 75.2% and 11.7%, respectively. The median Pearson correlation coefficient for this comparison was 0.85 and the corresponding scatter plot is shown in Fig. 5 (right part). Although there were no differences in performance and calibration between

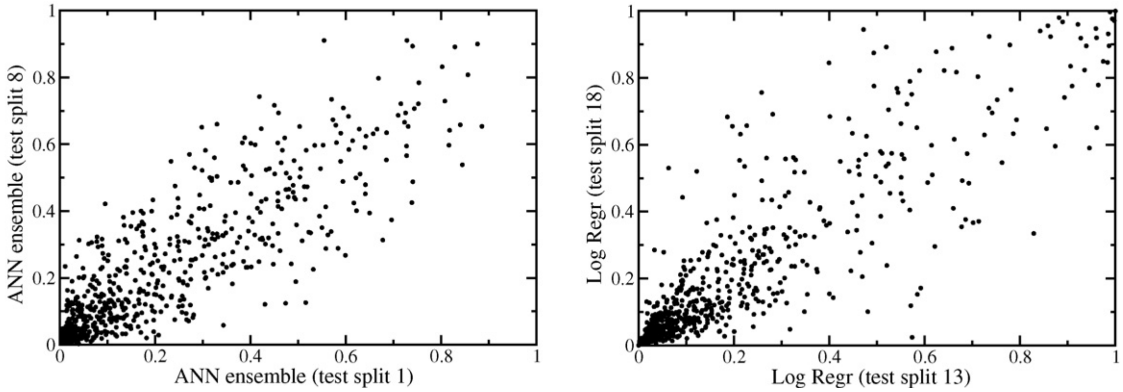

Figure 4 Intra-method scatter plots. The left figure shows the ANN cross-splitting ensemble ACS predictions for patients in test splits 1 and 8. The right figure are the corresponding ACS predictions for logistic regression model (test split 13 and 18). The ANN ensemble was trained on ECG data meanwhile the logistic regression model used both ECG and clinical data.

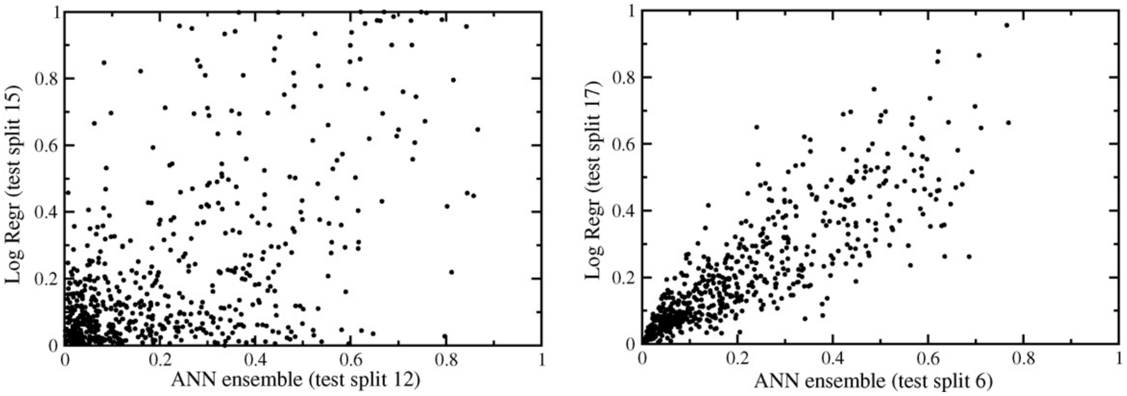

Figure 5 Inter-method scatter plots. The left figure shows ACS predictions for the ANN cross-splitting ensemble (ECG data) vs. the logistic regression model (all input features), using test split 12 and 15, respectively. The right figure corresponds to the bagging ensemble (clinical data) and the simplified logistic regression model, using test split 6 and 17.

these two models, there were still significant ACS prediction differences for specific patients. To further analyze the differences we looked at the 10 patients that had the largest ACS prediction differences in this scatter plot. The absolute differences ranged from 0.42 to 0.28. Four ACS patients was part of this subset and the ANN ensemble was correct in three cases. Among the remaining six non-ACS patients the ANN ensemble correctly classified four of them.

3.3. Comparing risk factors

For the logistic regression method one can easily compute odds ratios for each of the independent variables. Using odds ratios one can compare the different "predictor" variables. For the ANN ensemble one has to compute effective odds ratios because of the non-linearity in the model (see Section 2.5.1). Odds ratios for the logistic regression model and effective odds ratios for the ANN bagging ensemble are shown in Table 5. Both models were trained using only clinical data. For the ANN ensemble standard deviations were computed across patients. For both the logistic and the ANN ensemble model the odds ratios were computed using the full data set. For the ANN model this implied training an ANN ensemble on the full data set followed by the effective odds ratio calculation. For the logistic regression model odds ratios were calculated from the weights estimated using the full data set.

There was an overall good agreement between the odds ratios from the logistic regression model and the effective odds ratios obtained from the ANN bagging ensemble. Categorical factors with the largest odds ratios were symptom duration, angina pectoris, previous myocardial infarction and chest discomfort at presentation. It appears that the logistic regression model gave higher weight to

Table 5 Odds ratios and effective odds ratios for the logistic regression model and the ANN bagging ensemble

Variable	Logistic regression	ANN
Age	1.04	1.03 (0.01)
Gender		
Male	1.47	1.57 (0.42)
Diastolic blood pressure	1	0.99 (0.01)
Systolic blood pressure	1	1 (0.01)
Heart rate	1	1 (0.01)
Smoking status		
Current	1.59	1.37 (0.16)
Hypertension		
Yes	1.6	1.41 (0.18)
Diabetes		
Yes	1.15	1.07 (0.07)
Medication		
Yes	0.8	0.96 (0.13)
Angina pectoris		
Yes, ≤ 1 month	2.63	2.38 (0.58)
Yes, > 1 month	0.84	1.06 (0.3)
Congestive heart failure		
Yes	0.59	0.65 (0.1)
Chest discomfort at presentation		
Yes	2.14	2.2 (0.49)
Symptom duration		
0—6 h	5.12	3.79 (0.77)
7—12 h	3.8	2.67 (0.54)
13—24 h	1.33	1.02 (0.1)
Tachypnea		
Yes	1.01	1.15 (0.19)
Lung rales		
Yes	1.78	1.55 (0.15)
Previous myocardial infarction		
Yes, ≤ 6 months	3.19	2.94 (0.63)
Yes, > 6 months	1.86	1.97 (0.42)
Previous PTCA		
Yes	0.5	0.58 (0.11)
Previous CABG		
Yes	0.41	0.47 (0.11)

These models were trained using clinical data only. For the ANN ensemble the figures in parenthesis are standard deviations computed across patients.

"symptom duration" and that an "angina pectoris" event that occurred >1 month ago was not associated with a decrease in ACS risk, as in the logistic regression model. Neither of the models found the factors heart rate and diastolic and systolic blood pressure to be associated with any change of ACS risk.

4. Discussion

Part of the aim of this study was to construct a model for ACS prediction at the ED, only using data that are immediately available at presentation. The model was developed using data from chest pain patients at the ED of a university hospital and included clinical and ECG data. The best model was found to be an ANN cross-splitting ensemble, trained on ECG data only, with an area under the ROC curve of about 80%. The model was also well calibrated. There is a general consensus that ECG is one of the most important factors predicting ACS early at the ED. This is confirmed in this study since the best performance was obtained using only the ECG. Adding clinical information did not improve the performance for our study population. The obtained results did not confirm the high levels of ROC areas (>95%) found in other recent studies (e.g. [5,7,9]). One limiting factor in our study was the relatively small study population, however, this cannot be the only explanation. The prevalence of ACS was larger in the work of Kennedy and Harrison [7,9], ranging from 37% to 55% compared to a 21% prevalence of ACS in our study, which we believe is a more realistic number for an ordinary ED [32]. The prevalence of ACS in Baxt et al. [5] was as low as 16%. Furthermore, the presence of ST-elevation, ST-depression or T-wave inversion ECGs, in our population (see Table 2), was different compared to the cohorts of Kennedy and Harrison, where their training ACS (non-ACS) cases had 32% (1%) ST-elevation, 51% (1%) ST-depression and 44% (4%) T-wave inversion. It is apparent that ECG changes of this kind is very indicative of ACS and may therefore explain why ACS prediction was more difficult in our study population. Baxt et al. [5] obtained an ROC area of 90% with their ANN model, but this included a set of early chemical markers that was not part of our data, since we only included patient data immediately available at presentation. The ECG data used in our model was derived from measurements of the 12-lead ECGs and not from interpretations made by ED staff. The fact that our best model only used such ECG data is interesting since that would allow for a prediction model that is fully automatic without any manual intervention.

Part of this study was also to compare models based on ANN with logistic regression models. Since there are several aspects of how to measure the performance of a given prediction method, we used more than one measurement. The area under the ROC curve is a very popular performance measure in medical applications, but will of course not reveal differences for specific points along the ROC curve. Furthermore, the ROC curve is invariant under any transformation of the ACS predictions as long as the order of the individual ACS predictions is not changed. In a clinical setting however, it is important that the output value of the model can be interpreted as ACS predictions, i.e. we want a good calibration. One approach to measure the degree of calibration for the ACS predictions is the Hosmer—Lemeshow goodness-of-fit test [29]. Comparing models using the area under the ROC curve as performance measure we found an advantage using ANN ensembles compared to both single MLPs and logistic regression. The two different ensemble models tested, bagging and cross-splitting ensemble, obtained comparable ROC areas for the different sets of variables used. It is also apparent that using ensemble averaging increases the performance compared to the single MLP models. Using only clinical data, and no ECG data, there were no significant differences between logistic regression and ANN ensembles. Using only ECG data the performance was better for the ANN ensembles compared to the logistic regression model, indicating non-linear effects not captured by the linear model.

Comparing models using the Hosmer—Lemeshow test we found most ANN ensembles to be well calibrated with χ^2 values ranging from 11.6 to 14.5 with the corresponding *p*-value range of 0.17—0.07. For the logistic regression models the variation was larger ranging from 11.7 to 24.8 for the χ^2. Although the single MLP model using only ECG data obtained a larger ROC area compared to the corresponding logistic regression model, the calibration was much worse. It is obvious that there is no one-to-one correspondence between ROC area and calibration using the Hosmer—Lemeshow test, indicating that it is important to use both measurements for the final model selection. To continue the comparison between models we also looked at intra- and inter-method scatter plots, and the associated Pearson correlation coefficients, to reveal differences on a patient per patient basis. When comparing two models with the same ROC area and calibration statistics large differences for individual ACS predictions was found (see Fig. 5). An individual patient could be classified as having ACS using one method but with the other one the same patient would be at low risk.

The final choice of ACS prediction model, or even a combination of more than one model, has to be further analyzed and validated in properly designed prospective studies. A hybrid model consisting of both ANN ensembles and logistic regression models, each optimized using different input data, may turn out to be the overall best model.

4.1. Clinical implications

Because of possibly disastrous consequences of a missed case of ACS, the evaluation of patients with suspected ACS is very important. The quality of the current standard ED assessment is, however, insufficient. A large number of patients with suspected ACS are incorrectly hospitalized [2,1,33] and many patients with ACS are diagnosed only after lengthy (up to 12 h) observation, with a resulting delay in therapy and an impaired prognosis. At the same time, as many as 5% of those with ACS are erroneously sent home from the ED [34,32]. Thus, there is a great need for methods to improve ED evaluation. One such method is a decision support system based on ACS prediction models.

The best model developed in this study had a specificity of 41% at the sensitivity level of 95%. For our ACS prevalence of 21%, this corresponds to a positive predictive value of about 30% and a negative predictive value of 97%. The positive predictive value may seem low, but it is likely comparable to that of the ED physician's decision after current standard ED assessment, where some 70% of those admitted for suspected ACS prove not to have it [2,1,33]. We have been unable to find any published data on the positive predictive value of standard ED assessment for possible ACS.

Models for ACS prediction based on ECG and clinical characteristics can probably be applied in many different healthcare settings. For the present ACS prediction methods, it seems wise to exploit the reasonably high negative predictive value. Our models are thus probably best used as support for discharging a patient in healthcare settings where ACS prevalence is low, e.g. in primary care, in the initial ED triage or in telemedicine situations where information is limited. Adding the clinical judgment of a physician would probably increase the negative predictive value to close to 100%.

Whatever the use of our models, the limited number of variables imply a small need for manual input, and an increased likelihood that the model will actually be used in a busy environment. With the exception of the ACI-TIPI [4], the need for a time-consuming large input has been a weak point of several previous prediction models, e.g. [5], where up to 40 questions need to be answered before the model gives decision support.

4.2. Limitations and future work

The patients included in the present model were retrospectively collected and from one center only. Furthermore, the size of the collected dataset has an effect on the performance of the models and increasing the number of patients would probably lead to an increased performance. Before clinical implementation, the model clearly needs to be validated prospectively, preferably at multiple centers. To fully explore the use of ANN ensembles other techniques such as boosting or voting should be tested. Also the observed diversity between between logistic regression models and the ANN models could be utilized using a hybrid approach. The ECG representation using PCA may not be optimal and should be further investigated.

5. Conclusion

We have found that ANN ensembles, using ECG data only, can predict ACS at the ED with an area under the ROC curve of about 80%. No significant increase in performance was obtained adding clinical data available at presentation. Also, no significant differences were found between the bagging and the cross-splitting ensemble techniques. Comparing ANN ensembles with logistic regression models we found the former approach to be better in terms of ROC area and calibration assessments. Both ANN and logistic regression models showed intra-method variations, as a result of training the models with different parts of the study population. This variation was larger for the ANN ensemble models.

Acknowledgments

This study is in part supported by the Swedish Knowledge Foundation (2208/04-504). We are indebted to research nurse Ewa Mattson for collecting and arranging clinical data and for reviewing the final diagnose of each patient.

References

[1] Pope J, Ruthazer R, Beshansky J, Griffith J, Selker H. Clinical features of emergency department patients presenting with symptoms suggestive of acute cardiac ischemia: a multicenter study. J Thromb Thrombolys 1998;6:63—74.

[2] Ekelund U, Nilsson H-J, Frigyesi A, Torffvit O. Patients with suspected acute coronary syndrome in a university hospital emergency department: an observational study. BMC Emerg Med 2002;2:1—7.

[3] Goldman L, Cook EF, Johnson PA, Brand DA, Rouan GW, Lee TH. Prediction of the need for intensive care in patients who come to emergency departments with acute chest pain. N Engl J Med 1996;334(23):1498—504.

[4] Selker H, Beshansky J, Griffith J, Aufderheide T, Ballin D, Bernard S, et al. Use of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac ischemia. a multicenter, controlled clinical trial. Ann Intern Med 1998;129:845—55.

[5] Baxt W, Shofer F, Sites F, Hollander J. A neural network aid for the early diagnosis of cardiac ischemia in patients presenting to the emergency department with chest pain. Ann Emerg Med 2002;40:575—83.

[6] Xue J, Aufderheide T, Wright R, Klein J, Farrell R, Rowlandson I, et al. Added value of new acute coronary syndrome computer algorithm for interpretation of prehospital electrocardiograms. J Electrocardiol 2004;37:233—9.

[7] Harrison R, Kennedy R. Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Ann Emerg Med 2005;46: 431—9.

[8] Green M, Björk J, Hansen J, Ekelund U, Edenbrandt L, Ohlsson M. Detection of acute coronary syndromes in chest pain patients using neural network ensembles. In: Fonseca JM, editor. Proceedings of the second international conference on computational intelligence in medicine and healthcare. Lisbon, Portugal: IEE/IEEE; 2005. p. 182—7.

[9] Kennedy R, Harrison R. Identification of patients with evolving coronary syndromes by using statistical models with data from the time of presentation. Heart 2006;92:183—9.

[10] Baxt W. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Emerg Med 1991;115:843—8.

[11] Baxt W, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. Lancet 1996;347:12—5.

[12] Kennedy R, Burton A, Fraser H, McStay L, Harrison R. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: Derivation and evaluation of logistic regression models. Eur Heart J 1996;17:1181—91.

[13] Baxt W, Shofer F, Sites F, Hollander J. A neural computational aid to the diagnosis of acute myocardial infarction. Ann Emerg Med 2002;34:366—73.

[14] Hedén B, Öhlin H, Rittner R, Edenbrandt L. Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. Circulation 1997;96(6):1798—802.

[15] Ohlsson M, Öhlin H, Wallerstedt S, Edenbrandt L. Usefulness of serial electrocardiograms for diagnosis of acute myocardial infarction. Am J Cardiol 2001;88:478—81.

[16] Lisboa P, Ifeachor E, Szczepaniak P, editors. Artificial neural networks in biomedicine. London: Springer-Verlag; 2000.

[17] Hansen LK, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 1990;12:993—1001.

[18] Krogh A, Vedelsby J. Neural network ensembles, cross-validation, and active learning. In: Tesauro G, Touretzky D, Leen T, editors. Advances in neural information processing systems, vol. 2. San Mateo, CA: Morgan Kaufman; 1995 . p. 650—9.

[19] Opitz D, Maclin R. Popular ensemble methods: An empirical study. J Artif Intell Res 1999;11:169—98.

[20] Breiman L. Bagging predictors. Mach Learn 1996;24:123—40.

[21] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Raedt LD, Wrobel S, editors. Proceedings of the 22nd international conference on machine learning. Bonn, Germany: ACM Press; 2005.

[22] Ohlsson M, Öhlin H, Wallerstedt S, Edenbrandt L. Usefulness of serial electrocardiograms for diagnosis of acute myocardial infarction. Am J Cardiol 2001;88(5):478—81.

[23] Tunstall-Pedoe H, Kuulasmaa K, Amouyel P, Arveiler D, Rajakangas A, Pajak A. Myocardial infarction and coronary deaths in the world health organization monica project. Registration procedures, event rates, and case-fatality rates in 38 populations from 21 countries in four continents. Circulation 1994;90:583—612.

[24] Hanson SJ, Pratt LY. Comparing biases for minimal network construction with back-propagation. In: Touretzky DS, editor. Advances in neural information processing systems, vol. 1. Morgan Kaufmann; 1989. p. 177—85.

[25] Dietterich TG. Ensemble methods in machine learning. Lect Notes Comput Sci 2000;1857:1—15.

[26] West D, Mangiameli P, Rampal R, West V. Ensemble strategies for a medical diagnostic decision support system: a breast cancer diagnosis application. Eur J Oper Res 2005;162(2): 532—51.

[27] Hosmer D, Lemeshow S. Applied logistic regression New York: Wiley; 1989.

[28] Lippman R, Shahian D. Coronary artery bypass risk prediction using neural networks. Ann Thorac Surg 1997;63:1635—43.

[29] Hosmer DW, Hosmer T, le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16:965—80.

[30] Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982;143:29—36.

[31] Wehrens R, Putter H, Buydens L. The bootstrap: a tutorial. Chemometr Intell Lab Syst 2000;54:35—52.

[32] Pope J, Aufderheide T, Ruthazer R, et al. Missed diagnoses of acute cardiac ischemia in the emergency department. N Engl J Med 2000;342(16):1163—70.

[33] Karlson B, Herlitz J, Wiklund O, Richter A, Hjalmarson A. Early prediction of acute myocardial infarction from clinical history, examination and electrocardiogram in the emergency room. Am J Cardiol 1991;68:171—5.

[34] Lee T, Rouan G, Weisberg M, Brand D, Acampora D, Stasiulewicz C, et al. Clinical characteristics and natural history of patients with acute myocardial infarction sent home from the emergency room. Am J Cardiol 1987;60(4): 219—24.