Automatic Covariate Selection in Logistic Models for Chest Pain Diagnosis: A New Approach

Robert F Harrison, R Lee Kennedy [a,b]

[a] *Department of Automatic Control & Systems Engineering, The University of Sheffield, Sheffield, S1 3JD, UK*

[b] *School of Medicine, James Cook University, Townsville, QLD 4811, Australia*

Abstract

A newly established method for optimizing logistic models via a minorization-majorization procedure is applied to the problem of diagnosing acute coronary syndromes (ACS). The method provides a principled approach to the selection of covariates which would otherwise require the use of a suboptimal method owing to the size of the covariate set. A strategy for building models is proposed and two models optimized for performance and for simplicity are derived via ten-fold cross-validation. These models confirm that a relatively small set of covariates including clinical and electrocardiographic features can be used successfully in this task.

The performance of the models is comparable with previously published models using less principled selection methods. The models prove to be portable when tested on data gathered from three other sites. Whilst diagnostic accuracy and calibration diminishes slightly for these new settings, it remains satisfactory overall.

The prospect of building predictive models that are as simple as possible for a required level of performance is valuable if data-driven decision aids are to gain wide acceptance in the clinical situation owing to the need to minimize the time taken to gather and enter data at the bedside.

Key words: logistic regression, acute coronary syndromes, covariate selection, minorization-majorization algorithms
PACS:

Email address: r.f.harrison@sheffield.ac.uk, lee.kennedy@jcu.edu.au (Robert F Harrison, R Lee Kennedy).

1 Introduction

The selection of covariates for use in building predictive models has occupied statisticians and the machine learning community for many years. Exhaustive search is an obvious approach but is, of course, combinatorial in the number of covariates and soon becomes computationally impractical should this exceed about 20. Instead, forward selection adds new variables, one by one, choosing the one which most improves some measure of quality such as goodness-of-fit or predictive performance, at each stage. A popular variant on this performs a test at each step to determine if any existing variable can safely be removed without overly damaging performance. In contrast, backward elimination begins with all variables included and drops the least deleterious at each stage. Such a strategy is computationally more demanding than the forward algorithm. Neither approach is guaranteed to find a good subset, let alone one that is optimal. Indeed, even if both methods find identical sets, this may still be suboptimal. Branch and bound techniques provide an alternative, and are typically used to select the best subset of, or up to, a particular size. They have the advantage that an exhaustive search is conducted leading to an optimal solution. Of course, if the total number of subsets to be examined is large, exhaustive search becomes infeasible. A number of suboptimal selection methods is also discussed in [1]. Of these, [2] found that the so-called Sequential Forward Floating Search [3] method produced the best results, performing close to optimal and demanding lower computational resources than other methods. This method is a bottom-up search procedure, where the term *floating* identifies that the number of features changes dynamically, with one feature included and/or excluded, at each iteration. A fairly comprehensive treatment of the question of subset selection [4] describes a number of other methods at some length and touches on the more recent focus on "data-driven" approaches, whereby a modified (regularized) objective function that in some way penalizes the inclusion of terms that have low value is optimized e.g. via a maximum likelihood procedure. In particular, the "kernel machines" community has highlighted the role of regularization in this respect [5]. Here the sparsity inducing properties of certain regularizers is exploited to derive models that have low complexity whilst simultaneously achieving their goal of high predictive accuracy. This has worked well for dichotomies where a "yes/no" response is required but has been less successful when an estimate of probability is required or when the task is polychotomous.

By exploiting results from optimization theory known as optimization transfer or minorization-majorization (MM) algorithms [6, 7], it has recently been possible to extend sparsity-inducing regularization to the problem of multinomial logistic regression [8] resulting in an algorithm that is no more computationally demanding than conventional iterative schemes. This offers a principled alternative to the widely used, sub-optimal step-wise regression procedures

provided in a number of packages.

Our earlier work on the diagnosis of chest pain has demonstrated that relatively few covariates available at presentation can be combined through logistic regression or artificial neural networks to provide highly accurate diagnosis of acute myocardial infarction (AMI) [9] and acute coronary syndromes (ACS) [10, 11] but those models were constructed through *a priori* examination of the likelihood ratios of the various covariates. We recognize that this may not always lead to the best performing model and therefore we explore the more principled approach laid out here for the diagnosis of ACS.

1.1 Clinical Motivation

The diagnosis of acute coronary syndromes rests on clinical history, changes on the electrocardiogram (ECG) and cardiac marker protein data. Each of these evolves following presentation and is modified by treatment. Marker protein measurements provide definitive diagnostic and prognostic information but take several hours after the onset of symptoms to become positive. This has led to the development of protocols in chest pain units in many centres to manage patients in the early hours after the onset of symptoms, and before a definitive diagnosis can be made [12, 13]. A large proportion of patients who present to emergency departments with chest pain has non-cardiac diagnoses, and most of these would be, most appropriately, discharged directly home. In practice, a small, but significant, proportion of patients is sent home inappropriately [14] leading to potentially serious clinical errors and litigation. On the other hand, many relatively low-risk patients are inappropriately admitted to telemetry and high dependency units to rule out acute cardiac ischaemia [15]. In the centres used for that study, around 2% of patients were inappropriately discharged from emergency departments, while about 30% of patients presenting with acute chest pain were admitted with possible acute coronary syndrome but ultimately had the diagnosis ruled out.

Better use of clinical and ECG information available at presentation can improve identification of patients with evolving ACS. This has the potential to improve clinical care, since many triage and treatment decisions have to be made early, and could also optimize the use of resources, including chest pain units. Studies confirming that clinical, as well as ECG, factors are highly discriminatory for evolving ACS have strengthened research in this area recently [16, 17]. Various statistical and computer-based methods have been used to analyze clinical and ECG data from chest pain patients with a view to improving identification of high-risk patients at presentation. These methods include logistic regression [15, 9] classification trees [18, 19], and artificial neural networks (ANNs) [11, 20]. Each of these methods has advantages and

disadvantages although, suitably optimized, they can all provide accurate classification of low- and high-risk patients from data available at presentation.

In an earlier study [9] we found that a simple logistic model including only ECG data performed almost as well as a more extensive model incorporating clinical data items. The aim of that study was to develop a predictive model for AMI. Later we found that a logistic model to predict ACS placed additional importance on clinical factors [10]. The performance of this model was upheld when applied to subjects from other hospitals.

The authors of [15] have described a simple logistic regression model using mainly ECG data, the Acute Cardiac Ischaemia-Time Insensitive Predictive Instrument (ACI-TIPI), to identify patients with acute cardiac ischaemia. Use of ACI-TIPI in ten U.S. hospitals, increased the rate of discharge while decreasing inappropriate admission to high-dependency beds. Other studies have also demonstrated the potential for decision aids to improve admission and discharge practices for patients with acute chest pain [21, 22].

There has been little published on the specific topic of feature selection for the diagnosis of ACS/AMI. Baxt and colleagues [23] describe the development of an ANN for the diagnosis of AMI in the presence of missing data values. The technique reduces an initial set of 89 potential covariates to 40. The method used is not described in detail but is by "auditioning" subsets. This is unlikely to be exhaustive, given the combinatorial number of all possible subsets, hence suboptimal. Indeed, owing to the non-linear nature of the ANN, it is hard to see how else one might proceed other than by exhaustive testing. A later paper [20] widens the diagnosis to cardiac ischemia using the same sample and methodology. The results given there make use of cardiac marker data and deliver AUROC of 0.98 even with approximately 5% missing data – broadly in line with our results – although differences in study population and methodology make direct comparison impossible. In [24], the express problem of reducing the size of the covariate set (comprising ECG and clinical data) for ACS diagnosis is addressed via backward selection in logistic regression. Again direct comparison with our work is impossible but the results could yield a potentially useful system for guiding early discharge in low prevalence populations. In a later paper [25] backward selection was again used in a logistic model but covariate selection in the ANN was ad hoc, including or excluding ECG or clinical data. Unlike our earlier work [10, 11], these papers found substantial performance differences between the ANN and logistic models. For AMI screening [26] used methods of selecting the covariate set for an ANN based on forward/backward selection for logistic regression or via univariate statistical analysis but do not describe their inclusion/exclusion criteria. There is a danger inherent in using (generalised) linear methods for covariate selection for downstream use in ANNs - ANNs exploit, where they exist, non-linear relationships between covariates (e.g. interactions and higher-

order correlations) and reducing the covariate set in these ways excludes such relationships from the analysis. Paper [26] also used principal components analysis (PCA) as did [27]. Being linear, PCA also suffers from this problem and, while reducing the size of the input layer in ANNs still requires a full set of covariates to operate on. It is more properly a feature extraction technique rather than a covariate selection technique.

In order to gain widespread acceptance, a model should be easy to use in the emergency room, it should discriminate between low- and high-risk patients with a high degree of accuracy, be well calibrated, perform robustly with data from different institutions, and operate in a way that is clinically meaningful. To date, no algorithm has been described that fully satisfies these criteria.

The objective of this paper is to describe a method that permits the development of a simple system that goes some way towards meeting the above criteria. In particular, a principled approach to the reduction of the number of covariates required to make a prediction whilst maintaining an acceptable level of performance is introduced. This is integrated into a scheme for model development and validation leading to a system whose performance is comparable in terms of diagnostic ability and calibration to other published models but is less reliant on ad hoc selection criteria.

2 Methods and Theory

Logistic regression is a well-known technique for modelling the probability of an outcome, conditional on a particular set of evidence – the covariates – and has been widely applied in medical and clinical decision making. The outcome of the logistic regression optimization procedure is the maximum likelihood estimate of a set of coefficients that weight the individual covariates.

When the sample data are linearly separable [1] it is possible to make the likelihood function arbitrarily large so that one or more coefficients increases in magnitude without bound. This provides a motivation for penalizing the "size" of the coefficients and the use of quadratic (or weight-decay) regularization is well established in this context. Here, a term proportional to the sum of the squares of the coefficients is subtracted from the log-likelihood function so that maximization of the sum is a trade off between the size of the coefficients and the fit to the data. The uniqueness of the maximum is unaffected by the incorporation of the penalty and a simple modification of the the conventional iterative solution schemes is all that is required.

[1] They can be classified without error by the insertion of a hyper-planar decision boundary.

The introduction of a quadratic penalty into the optimization can be interpreted in the Bayesian framework as placing a prior distribution of Gaussian form on the values of the coefficients. This quadratic penalty solves the problem of separability and also suggests a method for selecting variables – those with relatively small coefficient magnitudes after optimization can be discarded – and this has been widely applied. While it is desirable that coefficients should be kept small if possible, the quadratic penalty tends rather to *discourage* large values and permits many small values to remain. This means that these may, collectively, contribute substantially to the result. If, instead of using a Gaussian penalty, a prior distribution with a sharp peak is used, this will have the effect of penalizing non-zero coefficients much more strongly so that the pay-off for setting small coefficients exactly to zero is relatively much greater. A log-likelihood penalty comprising the sum of absolute values of the coefficients, among others, has precisely this property and is the one used here. Again, the introduction of this penalty has no effect on the uniqueness of the maximum.

The "sparsity-inducing" property of this penalty is well-studied and has been used widely in the field of "kernel machines" – the Support Vector Machine and its variants in particular [5]. However, its introduction leads to a technical difficulty in the numerical optimization of the modified log-likelihood function that arises because the absolute value function has a discontinuous first derivative at the origin. The use of mathematical programming techniques for solving many problems in kernel machine learning overcomes this drawback for the class of problems mentioned above but these tend to require sophisticated algorithms and in many cases only provide binary decisions rather than probability estimates.

The use of the MM formulation overcomes this problem and permits the use of the absolute value penalty in conjunction with the log-likelihood for the logistic regression model. MM algorithms work by taking a difficult optimization problem and solving a nearby problem that happens to have the same solution. In the case of maximization, they rely on finding a surrogate function that *minorizes* the actual objective function – the value of the surrogate is everywhere less than or equal to the actual function and is tangent to it at the current estimate of the coefficient vector. Then, by *majorizing* the surrogate – finding its current maximum value – it can be shown that stepping to this value results in an increase in the value of the original objective function. A viable surrogate function for the log-likelihood function for dichotomous data was given in [28] and was extended to the multinomial situation in [6]. Iteration then results in convergence to the unique maximum. In [29] the MM method was applied to the least absolute deviation regression problem, leading to a viable surrogate for the absolute value penalty. By combining these ideas [8] provides an iterative algorithm for the maximization of the log-likelihood function with sparsity inducing penalty in the multinomial case – this is the method used here in

its binomial form. The resulting algorithm requires no more computational resource than the iteratively re-weighted least-squares method conventionally used for solving the logistic regression problem. The authors of [8] have provided a fairly comprehensive, downloadable package [30], however, it is easily programmed in the Matlab^{TM} environment [31], which is the approach adopted here.

Performance is measured in a number of ways, via the receiver operating characteristic (ROC) curve and its associated area (AUROC) [32] which provides a convenient, threshold-free method of assessing discriminatory ability, and by three other measures as recommended in [33, 34]:

Discrimination A measure of the ability of a classifier to select between diagnoses (here represented by the normalized Brier score). A "perfect" discriminator would have a normalized Brier score of unity.

Sharpness A measure of the confidence a classifier has in its predictions, rewarding "confident" diagnoses (close to one or zero) without regard to the quality of discrimination. Values close to unity are, again, desirable.

Reliability A measure of the difference between how well a classifier claims it can perform (sharpness) and how well it actually does perform (discrimination). The ideal value is zero with negative values indicating overconfidence and positive values, diffidence.

The *calibration* of the models – the match between predicted and observed proportions of patients with ACS – is then examined. We do this directly but also, as recommended in [35] to reveal whether or not any difference in the observed and expected proportions is related to their magnitude. Again, these measures of performance do not rely on a specific choice of threshold and permit an assessment of overall model quality.

Finally, because a classifier will be used to make a decision in practice, a strategy suited to the problem at hand must be chosen. Here we use a simple, "forced choice" scheme and choose a threshold such that sensitivity $\approx$ specificity – often referred to as the "optimal" threshold[2]. This is justified here since we seek only to summarize performance in a commonly understood way. In an operational situation it might be better to specify an acceptable level of specificity or positive predictive value (PPV) and set the threshold on that basis or, indeed, to opt for a risk-weighted scheme or introduce a reject option. We do not consider Bayes decision theory further here.

[2] Owing to its optimality in the case of equi-probable, Gaussian distributed data sources.

Clinical and ECG data were collected at presentation in the Emergency Departments of four participating United Kingdom teaching hospitals – The Royal Infirmary of Edinburgh (Hospital 1), The Western General Hospital, Edinburgh (Hospital 2), The Northern General Hospital, Sheffield (Hospital 3), and The Leicester Royal Infirmary (Hospital 4). Consecutive patients presenting with acute chest pain were recruited. All four hospitals are urban teaching hospitals. Hospitals 1 and 2 are in the same city and serve a population of just over 500,000. The Accident and Emergency Department of Hospital 1 receives around 90,000 patients per year. During the four-month period (August to December 1995) of data collection from this hospital, 4.2% of presentations were with acute, non-traumatic chest pain. Hospital 2, serving the same population as Hospital 1, receives medical emergencies through an acute assessment unit. It receives 25,000 patients per year, and during the period of data collection (February to August 1996), 10.1% of patients presented with chest pain. This high rate reflects the presence of a regional cardiac unit in Hospital 2, and the high proportion of patients diagnosed with ACS reflects the fact that many chest pain patients with less acute presentations in the city are seen in chest pain clinics and in a General Practice Assessment Unit. The demographics of chest pain presentations to Hospital 3 have been described previously [17] – the hospital serves a population of 530,000, and has 75,000 Emergency Department attendances per year, 4% of which are due to acute chest pain. Chest pain data from this hospital were collected over three months September to December in 1992. A small sample of patients was collected from Hospital 4.

Data from 1,253 consecutive patients presenting to Hospital 1 were used to derive our models. These were tested on prospectively collected data from 1,268 patients attending Hospital 2, 626 patients presenting to Hospital 3 and 152 patients presenting to Hospital 4.

The methods used for data collection have been described previously [9]. Training data for the models were obtained from 1,253 consecutive patients aged 18 years or over, presenting with non-traumatic chest pain to Hospital 1. The study included both patients who were admitted and those who were discharged. The attending doctors in the emergency department recorded clinical and ECG data on a purpose-designed proforma. Three researchers – Consultant Physician, Cardiology Registrar and Research Nurse, assigned the final diagnosis for all patients independently. This diagnosis made use of follow-up ECGs, cardiac markers, other investigations and clinical history obtained from the patient's follow-up notes. For patients discharged directly from the emergency department, or for those with incomplete follow-up, the patient or their General Practitioner was contacted for information about diagnosis or continuing symptoms one month after initial attendance. Further data to

test the models was obtained from the emergency medical units at Hospital 2 ($N = 1,268$), Hospital 3 ($N = 626$) and Hospital 4 ($N = 152$). The methods for data collection and diagnosis were as described above. Informed consent was obtained from all patients participating in the study, which was approved by the Medical Ethics Committees of the participating centres. In each hospital, patients were recruited 24 hours per day, and for seven days a week.

2.2 Methods of Measurement

All patients admitted to hospital had serial cardiac marker measurements in line with local protocols. The rate of missed diagnosis of ACS in those discharged was very low (less than 2%). Creatine kinase (CK) of greater than 180 U/L for women and 200 U/L for men was regarded as abnormal, as was CK-MB activity of greater than 5% of total CK activity, or a CK-MB mass of greater than 8 g/L. Troponin T or I was measured by standard radioimmunoassay (Boehringer) in patients admitted or regarded as being at high risk of ACS, and a value of greater than 0.1 g/L was regarded as abnormal [36]. ACS was diagnosed in all patients who had positive cardiac markers. Diagnosis of myocardial infarction was made on the basis of clinical history, serial ECGs and cardiac markers in line with current recommendations [37]. ST-segment elevation myocardial infarction (STEMI) was diagnosed when ST segment elevation exceeding 1 mm or pathological Q waves developed in two or more regional ECG leads. Non-ST-segment elevation myocardial infarction (non-STEMI) was diagnosed when positive cardiac markers were accompanied by changes (ST depression, T wave inversion) on sequential ECGs. Acute coronary syndrome without myocardial infarction was diagnosed when ECG changes not diagnostic of STEMI occurred in the absence of elevated markers, where elevated cardiac markers were not accompanied by ECG changes [38], where the patient had an unstable course necessitating acute cardiological intervention, when ST elevation exceeding 1.5 mm was present on stress testing, or when the patient suffered an adverse cardiac event (death, myocardial infarction, or need for urgent intervention) within 30 days of the initial event. Overall, stress testing was carried out on 15% of patients in the study.

2.3 Modelling Procedure

The training sample comprises all individuals from Hospital 1. Of the 40 covariates used in building logistic models for this task, 38 are nominal (binary-valued) variables listed in Table 1 and sorted in descending order of log-likelihood for ACS. The remaining pair are interval valued: patient age (years)

Covariate	LCL	LLR	UCL
Hypoperfusion	1.2	1.6	2.0
ST Depression	1.5	1.6	1.7
ST Elevation	1.4	1.5	1.6
New Q Waves	1.1	1.4	1.7
T Wave Inversion	1.1	1.1	1.1
Added Heart Sounds	-0.092	0.93	1.9
Crackles	0.84	0.88	0.91
Nausea/Vomiting	0.49	0.51	0.54
Right Arm Pain	0.35	0.36	0.38
Sweating	0.32	0.32	0.32
Diabetes	0.23	0.27	0.30
Worse With Movement	0.24	0.24	0.25
Left Arm Pain	0.21	0.21	0.21
Rhythm (AF/SVT)	0.14	0.20	0.26
Hypertension	0.18	0.19	0.21
Pain Described As "Tight"	0.18	0.18	0.18
Ex Smoker	0.15	0.16	0.17
Previous Angina	0.15	0.15	0.15
Retrosternal Chest Pain	0.14	0.14	0.15
Previous MI	0.11	0.12	0.12
Syncope	0.017	0.091	0.17
Hyperlipidaemia	0.0003	0.089	0.18
Family History IHD	0.058	0.068	0.077
Bundle Branch Block	0.013	0.051	0.09
Short Of Breath	0.027	0.031	0.035
Chest Pain Major Symptom	0.016	0.016	0.016
Pain Radiates To Back	-0.021	0.004	0.029
Sex	0.0	0.0015	0.003
Smoker	-0.037	-0.032	-0.027
Old MI On ECG	-0.13	-0.11	-0.085
Old Ischaemia On ECG	-0.40	-0.36	-0.32
Pain In Left Chest	-0.40	-0.38	-0.37
Pain In Right Chest	-0.45	-0.41	-0.37
Pain Described As "Sharp"	-0.58	-0.56	-0.54
Intermittent Pain	-0.91	-0.79	-0.68
Pain Affected By Posture	-1.1	-1.0	-0.97
Worse With Breathing	-1.7	-1.5	-1.3
Chest Wall Tenderness	–	$-\infty$	–

Table 1
Nominal covariates ranked by log-likelihood for ACS with lower and upper 95% confidence limits (LCL, UCL, respectively). The value $-\infty$ associated with "Chest Wall Tenderness" arises because, in this sample, no patient diagnosed with ACS presented with this symptom.

	ACS			not ACS		
	LCL	Mean	UCL	LCL	Mean	UCL
Age (years)	65.68	65.72	65.75	52.82	52.86	52.90
Duration (hours)	10.29	10.36	10.43	21.47	21.56	21.65

Table 2
Summary of statistics for interval-valued variables.

and duration of pain since onset of symptoms (hours) and are summarized in Table 2

To determine the optimal value of regularization parameter, ρ, a range of 50

Covariate	**M1** Coefficients	**M2** Coefficients
ST Depression	4.351	3.642
ST Elevation	3.939	3.389
T Wave Inversion	3.779	3.140
New Q Waves	1.304	0.239
Crackles	0.785	0.483
Nausea/Vomiting	0.593	0.147
Hypoperfusion	0.544	–
Smoker	0.461	–
Right Arm Pain	0.421	0.096
Sex	0.333	–
Pain Described As "Tight"	0.316	0.060
Sweating	0.253	0.208
Retrosternal Chest Pain	0.169	–
Ex Smoker	0.088	–
Left Arm Pain	0.033	–
Age (years)	0.031	0.012
Diabetes	0.026	–
Previous Angina	0.003	–
Old MI On ECG	-0.046	–
Hypertension	-0.080	–
Hyperlipidaemia	-0.097	–
Pain In Left Chest	-0.170	-0.298
Shortness Of Breath	-0.259	-0.067
Pain In Right Chest	-0.315	–
Pain Described As "Sharp"	-0.379	-0.404
Previous MI	-0.450	-0.094
Pain Affected By Posture	-0.489	-0.420
Old Ischaemia On ECG	-0.878	-0.411
Intermittent Pain	-1.148	-0.034
Pain Worse With Breathing	-1.339	-0.730
Chest Wall Tenderness	-1.509	–
Constant	-4.808	-2.706

Table 3
Selected covariates in **M1** and **M2** and their coefficients.

logarithmically-spaced values was used to estimate logistic models. At each value, 10-fold cross validation was undertaken to reduce bias in the estimate of predictive performance. At each value of ρ an ROC curve was constructed whose area was computed. The variation of AUROC with ρ is shown in the left panel of Figure 1. The right panel shows the three other measures of performance, discrimination, sharpness and reliability, computed simultaneously.

The task is now to determine the "best" single value of ρ. An obvious choice is the one corresponding to the maximum value of AUROC. However, it should be borne in mind that the higher the value of ρ, the larger the number of coefficients that will be zeroed thus eliminating the corresponding covariate. We use two strategies.

M1 – Largest AUROC Here the value of ρ corresponding to the maximum value of AUROC (AUROC = 0.9712) is selected. This is indicated by the leftmost vertical in Figure 1 and yields $\rho_1 = 1.3$. The measures of discrimination, sharpness and reliability are virtually unaffected by this choice.

M2 – Smallest Subset Here the user decides on an acceptable level of cross

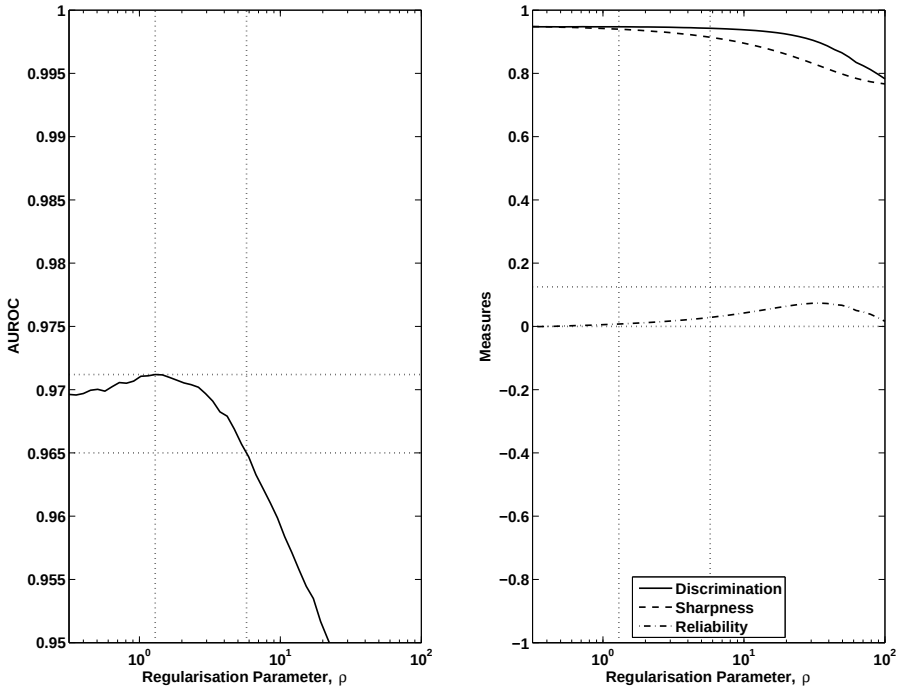

Fig. 1. Performance measures under 10-fold cross-validation: left panel – AUROC, right panel – discrimination, sharpness & reliability. The verticals indicate ρ_1 (left) and ρ_2 (right).

validated performance (AUROC = 0.9650 in this example) and then chooses the largest value of ρ that achieves it. This is indicated by the rightmost vertical in Figure 1 and yields $\rho_2 = 5.8$. Clearly, there is little degradation in the three other performance measures at this value and so it appears to be an acceptable choice.

Now the entire training sample is used to train the two models, **M1** and **M2** with ρ_1 and ρ_2, respectively. For illustration, in Figure 2, the coefficient values for **M2** are plotted in descending order of magnitude. It is clear that approximately one-half of them have negligible values. To quantify this, only those coefficients whose magnitude exceeds 0.5% of the coefficient with maximum magnitude are selected for the final model. The variables to which these correspond are listed for both models in Table 3 sorted from most positive to most negative. A constant is also retained in each model. There is no discernible difference between the performance of the models containing the negligible coefficients and the *reduced* models that omit them.

For final validation, the ROC curve is computed for each reduced model, **M1** and **M2** (see the top left panels of Figures 3 and 4, respectively) along with the other associated performance measures (Tables 4 and 5, respectively).

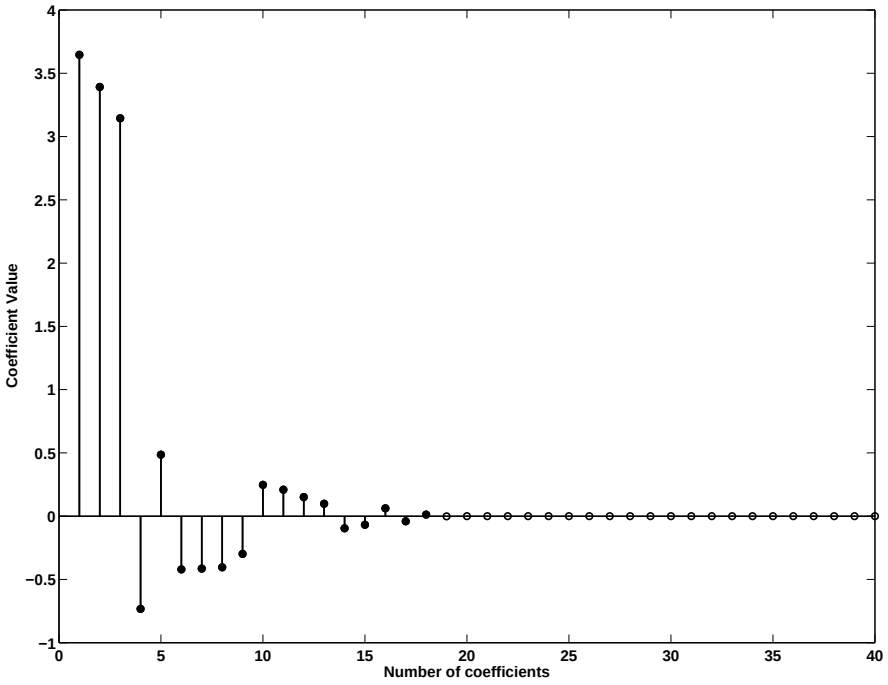

Fig. 2. Coefficient values for **M2** ranked in descending order of magnitude. Filled circles indicate the coefficients of the selected covariates.

It should be noted that the prior probability of ACS in each hospital is different from that of the training sample therefore each model has its outputs adjusted via a simple application of Bayes' Theorem to take this into account. In practice, local estimates of ACS prevalence will be known, e.g. from audit.

3 Results and Discussion

First we focus on the "threshold-free" performance measures which permit comparison of the models as a whole (provided their ROC curves do not substantially intersect). Figure 3 shows the ROC curves for each hospital for **M1** and figure 4 does likewise for **M2**. It is clear from this that there is no more than a 1% loss in AUROC in the transition to the smaller model. Indeed, this is also true for the conventional logistic regression estimate (not shown) whose AUROCs match those of **M1** to within 1%. In addition, the shapes of the curves remain virtually unchanged from model to model. Tables 4 and 5 reveal a similar picture in the measures of discrimination, sharpness and reliability, although here, sharpness and reliability can vary by 2–3%.

Figures 5 and 6 show the calibration, by decile, of the observed proportion of

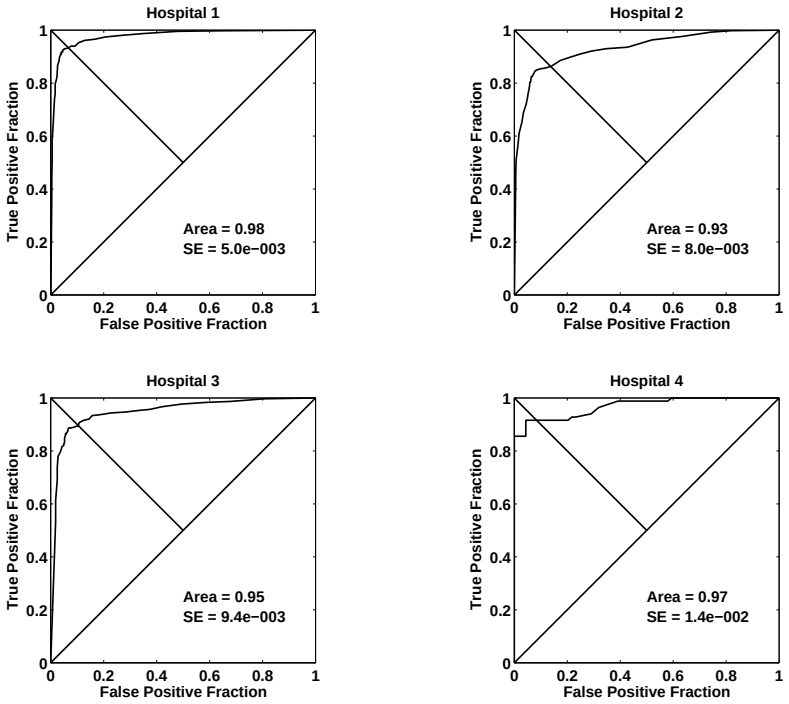

Fig. 3. ROC curves for **M1** applied to each hospital in the study.

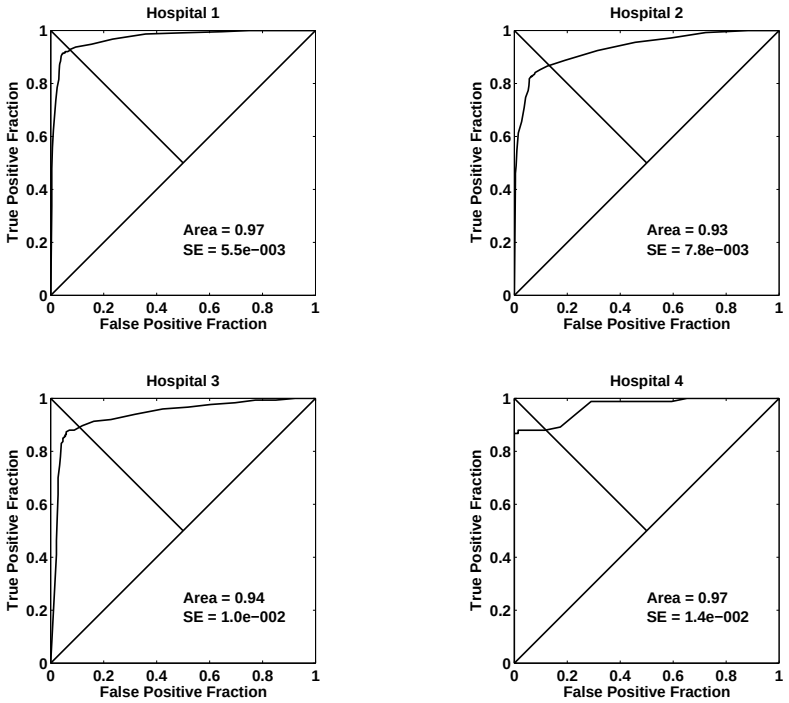

Fig. 4. ROC curves for **M2** applied to each hospital in the study.

	AUROC	SE	Discrimination	Sharpness	Reliability
Hospital 1	0.98	0.005	0.95	0.94	0.011
Hospital 2	0.93	0.008	0.91	0.93	-0.019
Hospital 3	0.95	0.009	0.92	0.93	-0.010
Hospital 4	0.97	0.014	0.94	0.91	0.024

Table 4
M1 threshold-free diagnostic performance measures.

	AUROC	SE	Discrimination	Sharpness	Reliability
Hospital 1	0.97	0.006	0.95	0.92	0.029
Hospital 2	0.93	0.008	0.91	0.90	0.006
Hospital 3	0.94	0.010	0.92	0.91	0.010
Hospital 4	0.97	0.014	0.93	0.88	0.050

Table 5
M2 threshold-free diagnostic performance measures.

	Lower 95% CL	Upper 95% CL
Hospital 1	0.073	0.097
Hospital 2	0.06	0.082
Hospital 3	0.037	0.062
Hospital 4	0.042	0.091

Table 6
M1 goodness of Fit – 95% CIs on residual mean-square. Small values (relative to unity) indicate a good fit.

the risk group versus the expected proportion predicted by the model outputs for each hospital. Ideally these points would fall upon the diagonal and [39] suggests a goodness-of-fit test based on residual sum-of-squares but for comparability across hospitals, we use the residual *mean*-square. A good fit is supported by small (compared with unity) values. Tables 6 and 7 provide 95% CLs for these values and support the hypothesis that both models provide a good fit.

An alternative recommendation for assessing calibration is to plot the difference of the values against their mean [35]. For good calibration, these would lie on the abscissa, any significant slope indicating that the calibration depends on the size of the values. Figures 7 and 8 therefore support the notion of good calibration since all but two values lie within the 95% CI (Hospitals 2 & 3) for **M1** and likewise for **M2** (Hospitals 1 & 3). In addition, the least-squares regression line is computed and plotted in each case. A regression line with

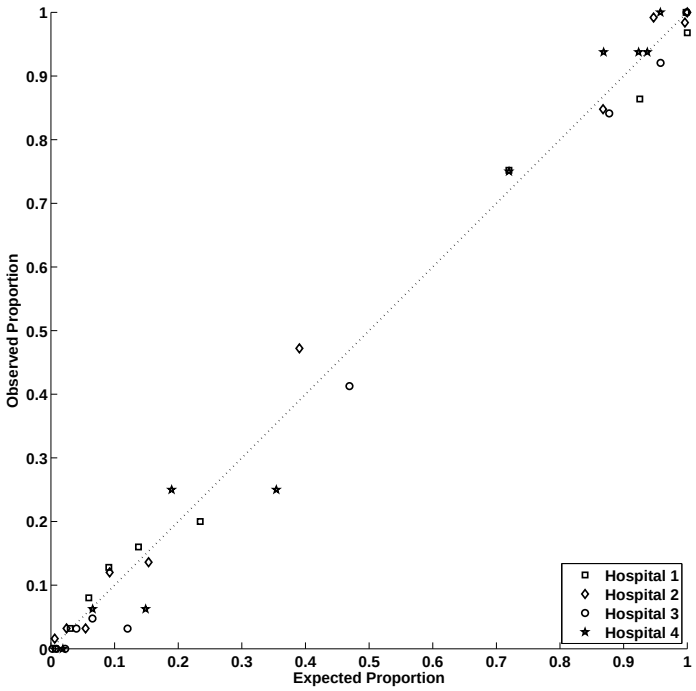

Fig. 5. Calibration data by percentile for **M1** applied to each hospital in the study. Ideally these would fall on the diagonal.

	Lower 95% CL	**Upper 95% CL**
Hospital 1	0.072	0.096
Hospital 2	0.065	0.087
Hospital 3	0.043	0.068
Hospital 4	0.044	0.09

Table 7
M2 goodness of Fit – 95% CIs on residual mean-square. Small values (relative to unity) indicate a good fit.

non-zero slope indicates a lack of calibration. However, the strength of the association may not be significant. A negative slope indicates an overestimate of small proportions and an underestimate of large ones and is characteristic of model overfitting.

For **M1** this is mildly evident for Hospital 4 while for Hospital 1, the opposite is true. Hospital 3 demonstrates a slight overall bias. For **M2** the situation apparently worsens: a negative slope for Hospitals 2–4 is now quite visible. Analysis of the 95% CLs for the correlation coefficients [3] (not shown) indicates

[3] Computed via 1000-fold bootstrap.

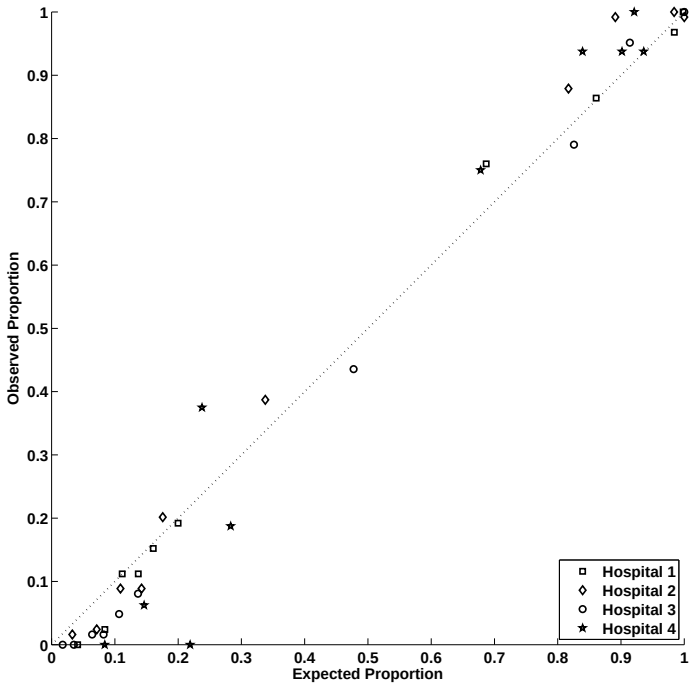

Fig. 6. Calibration data by percentile for **M2** applied to each hospital in the study. Ideally these would fall on the diagonal.

	Sensitivity	**Specificity**	**Accuracy**	**PPV**	**NPV**	**Threshold**
Hospital 1	0.94	0.93	0.94	0.90	0.96	0.23
Hospital 2	0.87	0.86	0.87	0.84	0.89	0.26
Hospital 3	0.90	0.90	0.90	0.89	0.90	0.34
Hospital 4	0.92	0.92	0.93	0.94	0.90	0.42

Table 8
M1 diagnostic performance at the "optimal" threshold (sensitivity = specificity).

that that the slopes are unlikely to differ from zero for **M1** but are inconclusive for **M2**. Since these are computed from only ten values, any judgement should be viewed with care.

Tables 8 and 9 summarize performance at the "optimal" threshold where sensitivity approximately matches specificity. Once again, the performance changes from the full model (not shown) to **M1** are no more than 1–2%. For Hospital 1–3 this is also reflected but for Hospital 4 there appears to be a degradation overall in the region of 5–7%. Notice, however, that the ROC curve is flat in this region probably caused by the relatively small sample for this hospital and so the value is not reliable.

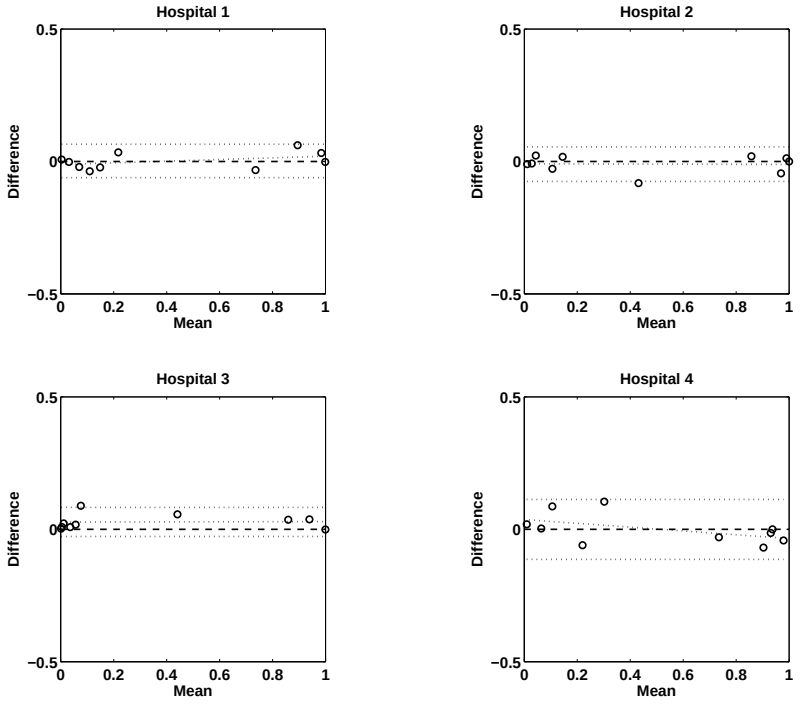

Fig. 7. Difference versus mean of calibration data for **M1** applied to each hospital in the study. Ideally these would lie on the abscissa.

	Sensitivity	**Specificity**	**Accuracy**	**PPV**	**NPV**	**Threshold**
Hospital 1	0.93	0.92	0.93	0.88	0.95	0.19
Hospital 2	0.86	0.88	0.84	0.80	0.90	0.19
Hospital 3	0.89	0.89	0.90	0.89	0.90	0.26
Hospital 4	0.86	0.88	0.84	0.87	0.85	0.30

Table 9
M2 diagnostic performance at the "optimal" threshold (sensitivity = specificity).

It is clear from the results above that the method of covariate selection is effective in building predictive logistic models for the diagnosis of ACS. The work confirms that both clinical and ECG data are important in making an accurate diagnosis and, furthermore, that these can be identified automatically according to a principled optimization procedure. The described methodology permits the designer a degree of freedom in determining whether accuracy or simplicity is of most value. Two models were derived to demonstrate this and, in terms of predictive ability, any differences in their performance was marginal. The models produced proved to be robust when used prospectively on data gathered from different settings and all appeared to be well calibrated. Poor calibration has been cited in the past as a potential disadvantage of e.g.

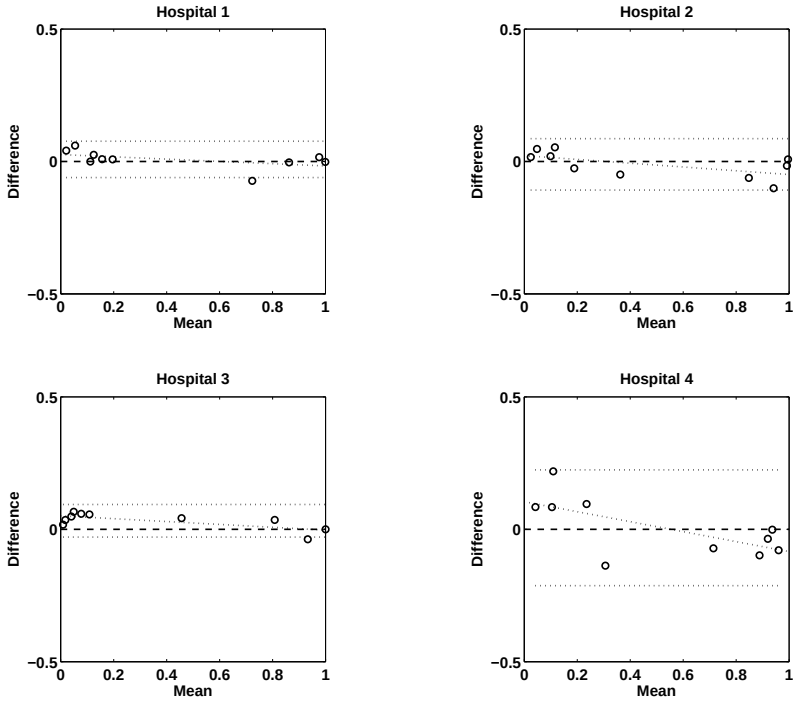

Fig. 8. Difference versus mean of calibration data for **M2** applied to each hospital in the study. Ideally these would lie on the abscissa.

neural network models for diagnosing acute cardiac ischaemia [40].

In an earlier study we applied a conventional logistic modelling approach and used the prior likelihood ratios of covariates as a method of term selection [10]. This worked well and the performance reported here is comparable to those models [4] . It is interesting that the terms selected by the present method match, quite closely, those of the earlier models. This is especially true of those of highest magnitude.

4 Conclusions and Recommendations

A model-building strategy making use of the MM optimization procedure has been proposed that permits optimization of a logistic model with a sparsity-inducing penalty. Two models are derived via ten-fold cross-validation: **M1** that optimizes predictive performance (as measured by AUROC) and **M2** that maintains a desired level of predictive performance using fewest covariates. Both models performed well against a variety of measures (diagnostic

[4] In that study interval-valued data were converted to a set of binary-valued design variables, so the comparison is not direct.

accuracy, calibration, sharpness and reliability) on data gathered from three other hospitals and had comparable performance to earlier published models – logistic and neural network – derived for this task. The results provide further confirmation that the automatic diagnosis of ACS can be made using a small number of covariates comprising both ECG and clinical information.

We have not included marker protein data in our study. Such data are difficult to collect systematically in samples of this size and, since they are used in the definition of the final diagnosis, incorporation into the set of explanatory variables would lead to misleading results. The potential benefit of using the model as proposed is not only that it would help rule in ACS quickly, but that it would assist with requests for investigations, including cardiac marker proteins. Many patients are discharged without investigation and one of the major benefits of this sort of decision aid would be to improve safety for early discharge. While cardiac markers would be useful for the model – indeed in [41] clinically useful, predictive values for diagnosis of AMI at two hours post admission were obtained using an ANN trained with serial measurements of either myoglobin or Troponin I – in clinical practice many patients do not have these measured. However, given the well-documented short- and long-term prognostic value of information derived from measurements of troponins and other proteins, future studies should examine how such measurements might be used alongside ECG and clinical data in developing a decision support system.

We are also unable to determine how the use of these models would influence diagnostic performance in practice. However, it is fair to say that any computational clinical decision aid should be easy to use to increase its chance of adoption. In particular, in the setting of a busy emergency department, it is important to minimize the time spent gathering and entering data at the bedside. We believe that the method proposed here offers a convincing, theoretically justified way of reducing model complexity, wherever possible, potentially enhancing useability of the final system by reducing the number of items to be input.

Acknowledgements

We are grateful to research staff who undertook collection and collation of clinical data Janine Robinson, Alex Burton, Louise McStay, Sue Mason and Taj Hassan. Hamish Fraser assisted with design of the study and with assigning final diagnoses. The study would not have been possible without consultants and junior doctors in the participating emergency departments filling in data proformas during already very busy working days.

References

[1] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 4–37.

[2] A. Jain, D. Zongker, Feature selection: Evaluation, application, and small sample performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 153–158.

[3] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (1994) 1119–1125.

[4] A. Miller, Subset Selection in Regression, 2nd Edition, Chapman & Hall, 2002.

[5] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.

[6] D. Böhning, Multinomial logistic regression algorithm, Annals of the Institute of Statistical Mathematics 44 (1992) 197–200.

[7] K. Lange, Optimization, Springer-Verlag, New York, 2004.

[8] B. Krishnapuram, L. Carin, M. A. Figueiredo, A. Hartemeink, Sparse multinomial logistic regression: fast algorithms and generalization bounds, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 957–968.

[9] R. Kennedy, A. Burton, H. Fraser, L. McStay, R. Harrison, Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models, European Heart Journal 17 (1996) 1181–1191.

[10] R. Kennedy, R. Harrison, Identification of patients with evolving coronary syndromes using statistical models with data from the time of presentation, Heart 92 (2006) 220–227.

[11] R. Harrison, R. Kennedy, Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation, Annals of Emergency Medicine 46 (2005) 431–439.

[12] M. Farkough, P. Smars, G. Reader, A. Zinmeister, R. Evans, T. Meloy, S. Kopecky, M. Allen, T. Allison, R. Gibbons, S. Gabriel, A clinical trial of a chest-pain observation unit for patients with unstable angina, New England Journal of Medicine 339 (1998) 1882–1888.

[13] F. Fesmire, A. Hughes, E. Fody, A. Jackson, C. Fesmire, M. Gilbert, P. Stout, J. Wojcik, D. Wharton, J. Creel, The erlanger chest pain evaluation protocol: A one-year experience with serial 12-lead ECG monitoring, two-hour delta serum marker measurements, and selective nuclear stress testing to identify and exclude acute coronary syndromes, Annals of Emergency Medicine 40 (2002) 584–594.

[14] J. Pope, T. Aufderheide, R. Ruthazer, R. Woolard, J. Feldman, J. Beshansky, J. Griffith, H. Selker, Missed diagnoses of acute cardiac ischaemia in the emergency department, New England Journal of Medicine 342 (2000) 1163–1170.

[15] H. Selker, J. Beshansky, J. Griffith, T. Aufderheide, D. Ballin, S. Bernard, S. Crespo, J. Feldman, S. Fish, W. Gibler, D. Kiez, R. McNutt, A. Moulton, J. Ornato, P. Podrid, J. Pope, D. Salem, M. Sayre, R. Woolard, Use of the Acute Cardiac Ischaemia Time Insensitive Predictive Instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms of acute cardiac ischaemia, Annals of Internal Medicine 129 (1998) 845–855.

[16] A. Panju, B. Hemmelgarn, G. Guyatt, D. Simel, The rational clinical examination. is this patient having a myocardial infarction?, Journal of the American Medical Association 280 (1998) 1256–1263.

[17] S. Goodacre, T. Locker, S. Campbell, How useful are clinical features in the diagnosis of acute undifferentiated chest pain?, Academic Emergency Medicine 9 (2002) 203–208.

[18] L. Goldman, F. Cook, P. Johnson, D. Brand, G. Rouan, T. Lee, Prediction of the need for intensive care in patients who come to emergency departments with acute chest pain, New England Journal of Medicine 334 (1996) 1498–1504.

[19] C. Tsien, H. Fraser, W. Long, R. Kennedy, Using classification tree and logistic regression methods to diagnose myocardial infarction, Medinfo 9 (1998) 493–497.

[20] W. Baxt, F. Shofer, F. Sites, J. Hollander, A neural network aid for the early diagnosis of cardiac ischaemia in patients presenting to the emergency department with chest pain, Annals of Emergency Medicine 40 (2002) 575–583.

[21] A. Qamar, C. McPherson, J. Babb, L. Bernstein, M. Werdmann, D. Yasick, S. Zarich, The Goldman algorithm revisited: Prospective evaluation of a computer-derived algorithm versus unaided physician judgement in suspected acute myocardial infarction, American Heart Journal 138 (2003) 705–709.

[22] J. Nicholl, R. Walls, L. Goldman, S. Pearson, H. Hartley, E. Antman, A critical pathway for management of patients with acute chest pain who are at low risk for myocardial ischaemia: Recommendations and potential impact, Annals of Internal Medicine 127 (1997) 996–1005.

[23] W. Baxt, F. Shofer, F. Sites, J. Hollander, A neural computational aid to the early diagnosis of acute myocardial infarction, Annals of Emergency Medicine 39 (2002) 366–373.

[24] J. Björk, J. Forberg, M. Ohlsson, L. Edenbrandt, H. Öhlin, U. Ekelund, A simple statistical model for prediction of acute coronary syndrome in chest pain patients in the emergency department, BMC Med Inform Decis Mak. 2006; 6: 28. 6 (2006) 28.

[25] M. Green, J. Björk, J. Forberg, U. Ekelund, L. Edenbrandt, M. Ohlsson, Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room, Artificial Intelligence in Medicine 38 (2006) 305–318.

[26] A. Bulgiba, M. Fisher, Using neural networks and just nine patient-

reportable factors of screen for AMI, Health Informatics Journal 12 (2006) 213–225.

[27] S. Olsson, M. Ohlsson, H. Öhlin, S. Dzaferagic, M. Nilsson, P. Sandkull, L. Edenbrandt, Decision support for the initial triage of patients with acute coronary syndromes, Clinical Physiology and Functional Imaging 26 (2006) 151–156.

[28] D. Böhning, B. Lindsay, Monotonicity of quadratic-approximation algorithms, Annals of the Institute of Statistical Mathematics 40 (1988) 641–663.

[29] K. Lange, D. Hunter, I. Yang, Optimization transfer using surrogate objective functions, Journal of Computational and Graphical Statistics 9 (2000) 1–59.

[30] B. Krishnapuram, L. Carin, M. A. Figueiredo, A. Hartemeink, SMLR: Sparse multinomial logistic regression, World Wide Web, `http://www.cs.duke.edu/~amink/software/smlr/` (2005).

[31] MathWorks, Matlab version 7.3, The MathWorks Inc., Natick, MA, 2005.

[32] J. Hanley, B. McNeil, The meaning and use of the area under a Receiver Operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[33] J. Hilden, J. Habbema, B. Bjerregaard, The measurement of performance in probabilistic diagnosis: II. trustworthiness of the exact values of the diagnostic probabilties, Methods of Information in Medicine 17 (1978) 227–237.

[34] J. Hilden, J. Habbema, B. Bjerregaard, The measurement of performance in probabilistic diagnosis: III. methods based on contiuous functions of the diagnostic probabilties, Methods of Information in Medicine (17) (1978) 238–246.

[35] J. Bland, D. Altman, Comparing methods of measurement: why plotting difference against standard method is misleading, Lancet 346 (1995) 1085–1087.

[36] M. Sayre, K. Kaufmann, I. Chen, M. Sperling, R. Sidman, D. Dierks, T. Liu, W. Gibler, Measurement of cardiac troponin T is an effective method for predicting complications among emergency department patients with chest pain, Annals of Emergency Medicine 31 (1998) 539–549.

[37] The Joint European Society of Cardiology/American College of Cardiology Committee, Myocardial infarction redefined - a consensus document of the Joint European Society of Cardiology/ American College of Cardiology Committee for the Redefinition of Myocardial Infarction, European Heart Journal 36 (2000) 959–969.

[38] P. Collinson, P. Stubbs, A. Kessler, Multicentre evaluation of the diagnostic value of cardiac troponin T, CK-MB mass, and myoglobin for assessing patients with suspected acute coronary syndromes in routine practice, Heart 89 (2003) 280–286.

[39] D. Hosmer, T. Lemeshow, S. le Cessie, A comparison of goodness-of-fit tests for the logistic regression model, Statistics in Medicine 16 (1997) 965–980.

[40] H. Selker, J. Griffith, S. Patil, W. Long, R. D'Agostino, A comparison of performance of mathematical predictive methods for medical diagnosis: Identifying acute cardiac ischaemia among emergency department patients, Journal of Investigative Medicine 43 (1995) 468–476.

[41] K. Eggers, J. Ellenius, M. Dellborg, T. Groth, J. Oldgren, E. Swahn, B. Lindahl, Artificial neural network algorithms for early diagnosis of acute myocardial infarction and prediction of infarct size in chest pain patients, International Journal of Cardiology 114 (2007) 366–374.