# Intelligent computer reporting 'lack of experience': a confidence measure for decision support systems

**H. Holst[1], M. Ohlsson[2], C. Peterson[2] and L. Edenbrandt[1]**

Departments of [1]Clinical Physiology and [2]Theoretical Physics, Lund University, Lund, Sweden
Received 28 June 1997; accepted 8 October 1997
Correspondence: H. Holst, Department of Clinical Physiology, University Hospital, S-221 85 Lund, Sweden

## Summary

The purpose of this study was to explore the feasibility of developing artificial neural networks that are able to provide confidence measures for their diagnostic advice.

Computer-aided decision making can improve physician performance, but many physicians hesitate to use these 'black boxes'. If we are to rely upon decision support systems for such tasks as medical diagnosis it is essential that the computers indicate when the advice given is based on experience, i.e. give a confidence measure.

An artificial neural network was trained to diagnose healed anterior myocardial infarction and to indicate 'lack of experience' when test electrocardiograms were different from the electrocardiograms of the training set. A database of 1249 electrocardiograms from patients who had undergone cardiac catheterization was used to train and test the neural network. Thereafter, the ability of the network to indicate 'lack of experience' was assessed using 100 left bundle branch block electrocardiograms, an electrocardiographic pattern that was excluded from the training set.

The network indicated that 83% of the left bundle branch block electrocardiograms and 1% of the test electrocardiograms from catheterized patients were different from the electrocardiograms of the training set. All but one of the left bundle branch block electrocardiograms would otherwise be falsely classified as anterior myocardial infarction by the network.

Artificial neural networks can be trained to indicate 'lack of experience', and this ability increases the possibility for neural networks to be accepted as reliable decision support systems in clinical practice.

*Keywords:* artificial intelligence, computer-assisted, diagnosis; electrocardiography, myocardial infarction.

## Introduction

Computers can be used to support physicians in situations where more experienced colleagues are not present. It has been shown that computer-aided decision making can improve physician performance (Johnston *et al*., 1994). The most widely used decision support systems in the medical field are probably computer programs for interpretation of electrocardiograms (ECGs). An estimated 300 million ECGs are recorded and interpreted by computerized electrocardiographs each year. It has been demonstrated that the best of these interpretation programs perform almost as well as human experts (Willems *et al*., 1991). Computer-aided decision making has been used in many other applications, for example image interpretation (Scott, 1993; Kahn, 1994), and in the clinical laboratory (Place *et al*., 1994). The standard methods for classifying clinical data have been knowledge-based systems and statistical methods such as linear discriminant analysis. Over the past 5 years a third method, artificial neural networks, has been introduced in the medical field (Baxt, 1995; Cross,

1995; Dybowski & Gant, 1995; Hedén *et al.*, 1995, 1996, 1997). Recently, neural networks were implemented in computerized electrocardiographs for the diagnosis of myocardial infarction.

Even though many decision support systems are highly accurate, few of them have been widely used. Many physicians hesitate to use these 'black boxes' because the reasoning behind the computer judgements is not transparent. In contrast, the same physicians would probably be willing to rely on the advice offered by a colleague even if he or she was unable to give the reason behind this advice. However, a colleague can report whether the advice given is based on experience from similar situations or not. If we are to rely upon computer-aided decision support systems for such tasks as medical diagnosis it is essential that the computers indicate when the advice given is based on experience, i.e. give a confidence measure.

Yang *et al.* (1994) demonstrated this problem in an electrocardiographic study. They trained a neural network to diagnose inferior myocardial infarction with ECGs recorded on normal subjects and patients with inferior myocardial infarction. Thereafter, the network was tested on ECGs with patterns of left ventricular hypertrophy and left ventricular strain. More than 25% of these ECGs were falsely classified as infarction. An indication that the neural network did not have experience in these patterns would have been more appropriate.

The purpose of the present work was to explore the feasibility of developing artificial neural networks that are able to provide confidence measures for their diagnostic advice, i.e. to indicate when a case differs from those of the training set. The method, which is generally applicable, was applied in an electrocardiographic classification task.

A neural network was trained to diagnose anterior myocardial infarction using a database of ECGs recorded on patients who had undergone cardiac catheterization. It is generally appreciated that the diagnosis of healed anterior myocardial infarction from a single ECG is impossible or at least very difficult in the presence of complete left bundle branch block, and ECGs with this pattern are therefore often excluded in studies concerning myocardial infarction and ECG. In this study, left bundle branch block ECGs were excluded from the group used to train the neural network. The network was trained not only to diagnose anterior myocardial infarction and normal ECGs, but also to find all types of ECG patterns that were not represented in the training set. After the training process, the network was tested on a group of left bundle branch block ECGs. These ECGs were used as one small sample of all possible strange ECG patterns that are found in clinical practice but not in a database used to develop a decision support system.

## Materials and methods

### Material

A total of 1249 ECGs from the Pahlm, Haisty, Bowman Gray School of Medicine Data Base (Pahlm *et al.*, 1991) recorded on patients who had undergone diagnostic cardiac catheterization were studied. The ECGs were used to train and test an artificial neural network for the diagnosis of anterior myocardial infarction. Anterior myocardial infarction was defined by presence of ≥75% diameter stenosis of the left main coronary artery, the left anterior descending artery or its major diagonals, and akinesia or dyskinesia of the anterior–superior wall on the right anterior oblique ventriculogram. Patients with normal coronary arteries, normal contrast left ventriculogram, no evidence of valve dysfunction or congenital heart disease, ejection fraction ≥50%, and an overall study evaluation of 'normal' were classified as 'cath normal'. Inferior myocardial infarction was defined by the presence of ≥75% diameter stenosis of the right coronary artery and akinesia or dyskinesia of the inferior wall on the right anterior oblique ventriculogram. Multiple myocardial infarction was defined by the presence of both anterior and inferior myocardial infarction.

The ECGs recorded on patients with anterior or multiple myocardial infarction were denoted the infarct group. Cath normals and patients with single inferior myocardial infarction constituted the control group. Approximately two-thirds of the ECGs in each of the diagnostic groups were randomly allocated to a training set and the remaining ECGs constituted a test set. The numbers of ECGs in the different groups are presented in Table 1.

One hundred ECGs with left bundle branch block were selected from the data bank of the Department of Clinical Physiology, University Hospital, Lund, by

**Table 1** Study population

| | Training set | Test set | Total |
|---|---|---|---|
| Infarct group | | | |
| Anterior infarction | 182 | 90 | 272 |
| Multiple infarction | 94 | 48 | 142 |
| Control group | | | |
| Inferior infarction | 238 | 118 | 356 |
| Cath normal | 320 | 159 | 479 |
| Total number of cath patients | 834 | 415 | 1249 |
| Artificial cases | 834 | 0 | 834 |
| Left bundle branch block | 0 | 100 | 100 |
| Total number of cases | 1668 | 515 | 2183 |

an experienced electrocardiographer. Left bundle branch block was defined by the presence of QRS duration of 0.12 s or more, left-sided precordial leads with absent Q waves and broad and notched or slurred R waves (Willems *et al.*, 1985).

The 12-lead ECGs were recorded using computerized electrocardiographs. The recording technique of the different electrocardiographs was in accordance with AHA specifications. The frequency range was 0.05–100 Hz and noise reduction was made by time coherent averaging. Averaged complexes were transferred to a computer and stored for further analysis. Measurements of amplitudes and durations of the electrocardiographic complexes were performed using custom software. The following 27 automated QRS and ST–T measurements were used as inputs to the artificial neural networks: nine measurements from each of leads V2, V3 and V4, namely Q, R and S duration and amplitude, three amplitudes within the ST–T segment regularly spaced between, but not including, ST junction and end of T. It should be stressed that in all types of automated ECG interpretation only a selection of leads and measurements are used. In contrast, physicians can look at the complete ECG complexes in all 12 leads. Conventional ECG interpretation programs (Macfarlane & Lawrie, 1989) and artificial neural networks (Hedén *et al.*, 1994) use leads V2, V3 and V4 for the diagnosis of anterior myocardial infarction. Consequently, the programs and networks are not influenced by the ECG appearance in other leads, for example broad and slurred R waves in leads V5 and V6 are not taken into account.

### Artificial neural networks

Artificial neural networks with a multilayer perceptron architecture (Rumelhart & McClelland, 1986) were used. A more general description of neural networks can be found elsewhere (Cross *et al.*, 1995). The neural networks consisted of 27 units in the input layer, five units in the hidden layer and three units in the output layer. The first unit in the output layer represents infarct ECGs, the second control ECGs and the third indicates 'lack of experience'. Each output unit gives a value between 0 and 1. The number of input units was equal to the number of input variables. During a training process, the connection weights between the units were adjusted using the back-propagation algorithm with Langevin updating (Rögnvaldsson, 1994). The learning rate was decreased geometrically every epoch from a start value of 0.5 to an end value of 0.1. The momentum was set to 0.7. Training was terminated at an error of 0.29 in order to avoid 'overtraining'. This error threshold was decided using a threefold cross-validation procedure. The performance of the neural network was studied using the test set only. The left bundle branch block ECGs were only used in the test set. All calculations were done using the JETNET 3.0 package (Peterson *et al.*, 1994).

### Confidence measure

A confidence measure, which assesses whether or not the neural network has experience in similar cases from the training process, was calculated (Roberts *et al.*, 1994). The different steps of this method are illustrated with a constructed example in which a neural network is trained to separate normal cases from infarcts (Fig. 1). In this example, only two input variables are used such that the data can easily be visualized (Fig. 1a). However, the method is also applicable in a multidimensional data space, which is common in clinical problems (27 variables were used in the electrocardiographic classification task described below).

Step 1: Each variable in the training set is normalized such that the mean of all examples is zero with a unit variance.

Step 2: A set of artificial cases is constructed by generating a random sequence of variables whose

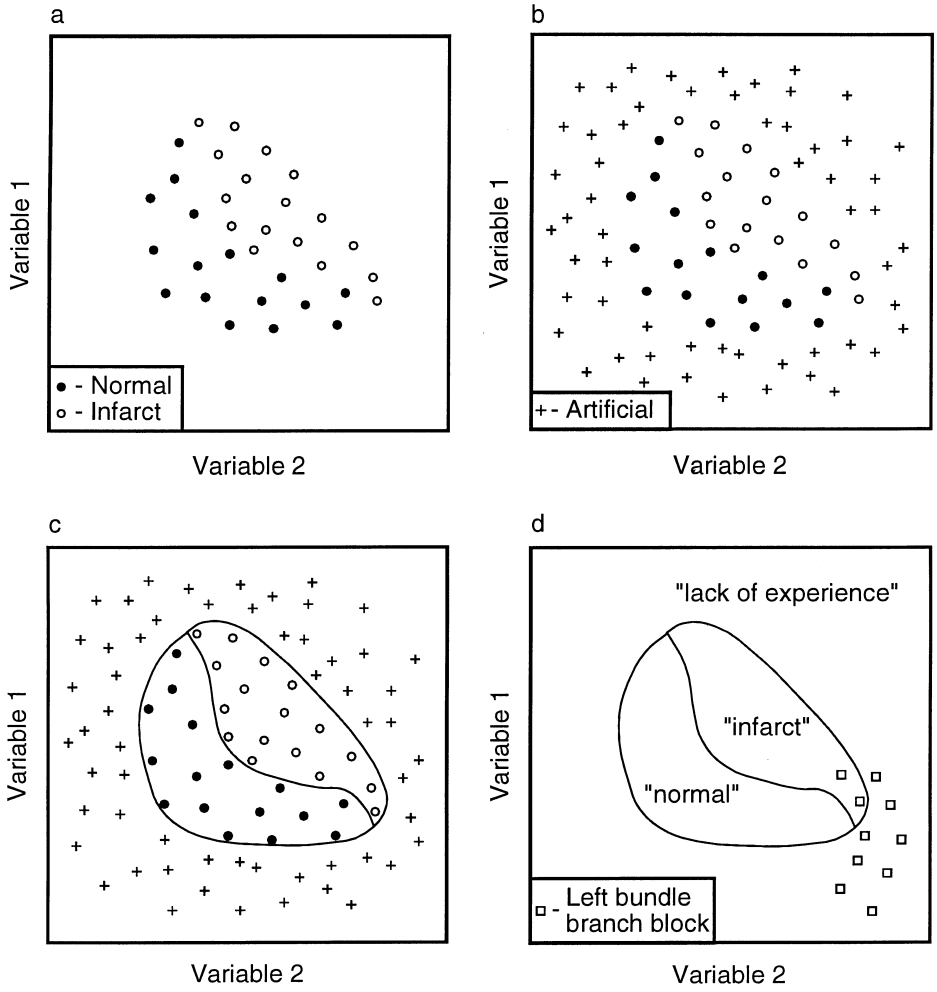


**Figure 1** The method for calculating a confidence measure is illustrated with an example. (a) Scatterplot showing the distribution of two groups (normal and infarct). (b) Artificial cases are added and those whose nearest neighbour is a normal or an infarct case are removed. (c) Classification boundaries of a neural network separating infarct, normal and the surrounding artificial cases. (d) Cases of a third class (left bundle branch block) are tested. The network will report 'lack of experience' in most of these cases.

values have zero mean and a variance, $k$, greater than unity. The artificial cases are distributed in a larger region than the real cases and this region includes the region of the real cases.

Step 3: Artificial cases which are in the region of the real cases could impair the network's ability to classify the real cases as normal or infarction. Therefore, such artificial cases are removed. All artificial cases whose nearest neighbour is a real case are removed using an iterative method. The remaining artificial cases which surround the real cases are added to the training set (Fig. 1b).

Step 4: The training set consists of three groups: normal cases, infarcts and artificial cases. A neural network with three output units is trained to separate the groups. After the training process, the network separates the three groups according to the boundaries in Fig. 1c.

Step 5: The performance of the network is studied using the test set. A low output from the third unit indicates that the test case studied is similar to real cases in the training set and vice versa. Test cases with an output value from the third unit less than a predefined threshold are classified as normal or

infarct, depending on the output values of the first and second unit. For test cases with an output larger than the threshold, a statement such as 'lack of experience' is justified (Fig. 1d).

In the electrocardiographic study, the training set consisted of 834 ECGs. After the normalization procedure (step 1) the same number of artificial cases was constructed using $k = 3$ (step 2). After removal of artificial cases that had a real case as closest neighbour, approximately 120 artificial cases remained (step 3). Step 2 and 3 were repeated until a total of 834 artificial cases could be added to the training set. The neural network was trained using the 834 real cases and the 834 artificial cases (step 4). Thereafter, the test set, consisting of 415 infarct and control ECGs and 100 left bundle branch block ECGs, was analysed (step 5).

## Results

The network output value from the third unit (lack of experience) was lower than 0.23 in 99% of the infarct and control ECGs in the test set. These ECGs were classified as infarct or control with a sensitivity of 73% and a specificity of 95%. Leads V2–V4 of the remaining four ECGs (1%) are presented in Fig. 2. These ECG patterns with (a) tall T waves, (b) broad R waves, (c) ST depression in lead V2 in combination with ST elevation in lead V4 and (d) tall R waves are not commonly found. The output value from the third unit was larger than 0.23 in 83% (83/100) of the left bundle branch block ECGs. The network outputs in these 87(4 + 83) ECGs can be translated into 'lack of experience'.

The network classified the remaining 17% of the left bundle branch block ECGs as anterior infarction, indicating that similar cases were part of the training set. One of these ECGs is presented in Fig. 3. An infarct ECG from the training set is shown in the same figure. The similarities between the ECGs in leads V2–V4 indicate why the network classified the test ECG as anterior infarct. The typical left bundle branch block pattern in this example is found in leads V5–V6, which were not presented to the network. As mentioned above, these leads are not used for the diagnosis of anterior myocardial infarction by conventional interpretation programs or by artificial neural networks.

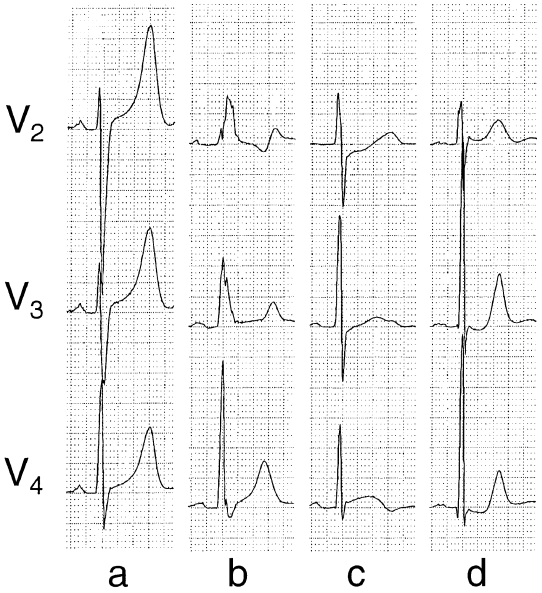


**Figure 2** Neural network reported 'lack of experience' in these four test electrocardiograms. Note the patterns of (a) tall T waves, (b) broad R waves, (c) ST depression in lead V2 in combination with ST elevation in lead V4 and (d) tall R waves, which are not commonly found.

A lower threshold applied to the third unit will result in an increased number of ECGs reported as 'lack of experience' and vice versa. All infarct and control ECGs in the test set have an output from the third unit lower than 0.46, a value which was exceeded in 48% of the left bundle branch block ECGs. Conversely, a threshold of 0.17 resulted in 2% infarct and control ECGs and 88% left bundle branch block ECGs reported as 'lack of experience'. The terms sensitivity and specificity are not used to describe these results because the statement 'lack of experience' is not strictly false positive, for example for the ECGs in Fig. 2. The confidence measure, i.e. the output value of the third unit, is only used to assess the similarities between a test case and those of the training set. Different thresholds could be applied to the third unit depending on what is regarded as an acceptable level of basis for the computer-based advice in the particular clinical situation.

If the confidence measure had not been taken into account, all test ECGs would have been classified as infarct or not. In the left bundle branch block group, 99 of the 100 ECGs would have been classified as anterior myocardial infarction.

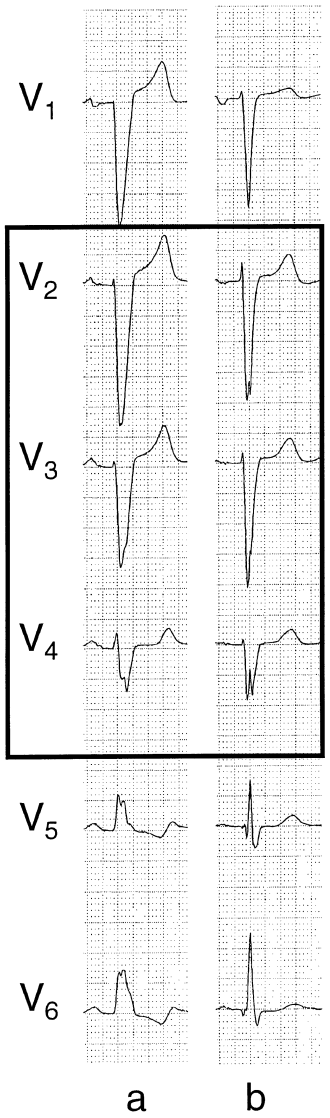


**Figure 3** The network classified this (a) left bundle branch block electrocardiogram (ECG) as anterior myocardial infarction. There are great similarities with (b), the infarct ECG of the training set in the leads V2–V4. The similarities between the ECGs give an indication of why the network made this classification. The typical pattern of left bundle branch block is found in leads V5–V6, leads which were not presented to the network.

## Discussion

### Main findings

Left bundle branch block could easily be detected using parameters such as QRS duration and R duration in leads V5, V6, aVL and I. However, the purpose of this work was not to diagnose left bundle branch block, but rather to explore a method that can be used to indicate when a case differs from those of a training set irrespective of what type of unseen pattern is present. The left bundle branch block ECGs represented a group likely to have a high frequency of such cases. The network indicated a lack of experience in 83% of the left bundle branch block ECGs and it could thereby refrain from classifying these ECGs as anterior infarct. A conventional decision support system would have made a classification without any indications that it has little or no basis for coming to its conclusion. The advantage of the method is that not all types of patterns which could be presented to the network when used in clinical practice need to be collected and presented to a network in a training session. This is an impossible task.

### Conventional discriminant analysis

The advantage of the neural network method presented in this study over conventional discriminant analysis is illustrated in Fig. 4. Two parameters are used to discriminate between normal subjects and patients with infarct. The groups could be separated with a deterministic approach (Fig. 4a), for example:

> IF parameter 1 is greater than limit I and parameter 2 is greater than limit II THEN report infarct.

All individuals in the upper right corner of the diagram would be reported as infarct. Similarly, the groups could be separated using a traditional statistical techniques, for example linear discriminant analysis (Fig. 4b), or an artificial neural network (Fig. 4c). The three methods are all developed to separate normal subjects from patients with infarct in the region where training data are available. The boundaries are developed in different ways, but a common feature is that an extrapolation outside this region not could be justified. The method presented in this study separates different groups in the region where training data are available, but also indicates when a case is different from those of the training data (Fig. 4d). The region where training data are available is delineated with a closed curve that is impossible to

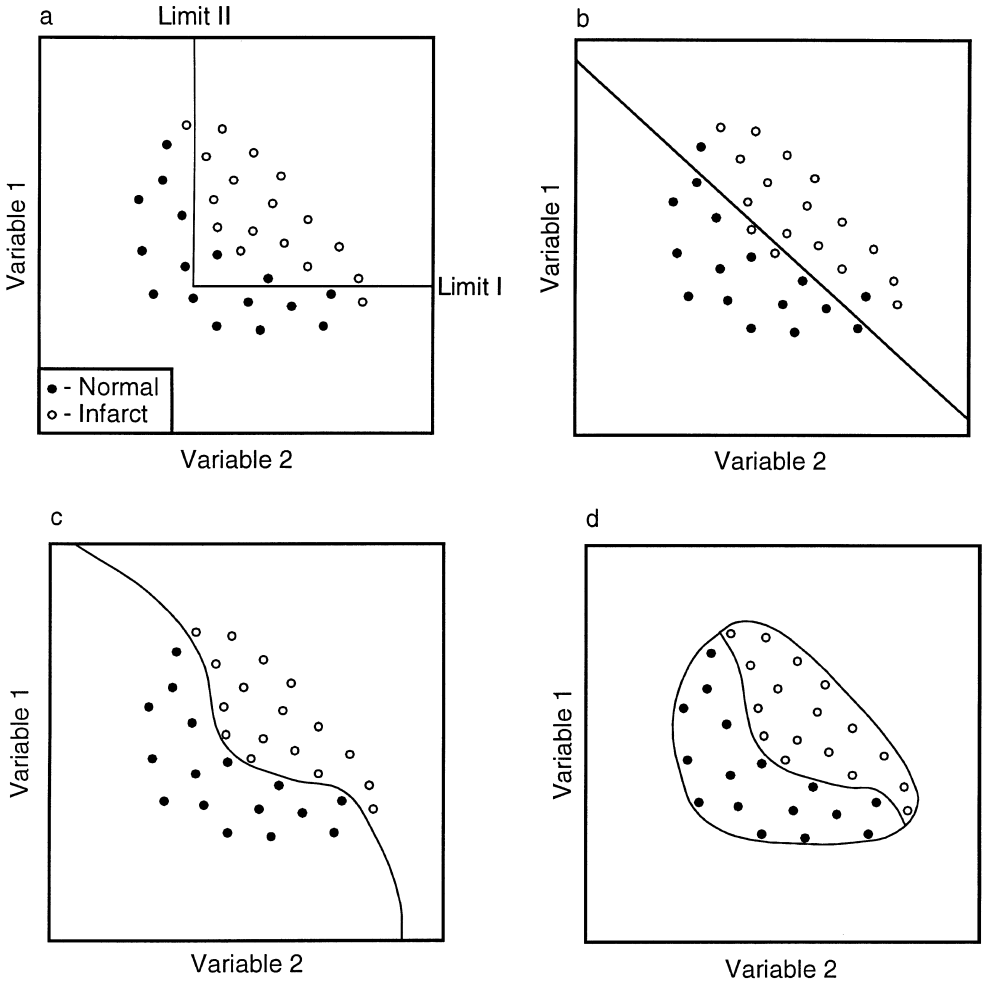


**Figure 4** Scatterplots showing the distribution of normals and infarcts in a training set. The groups are separated using (a) a decision tree method, (b) linear discriminant analysis and (c) a conventional neural network. Note that the three methods produce boundaries which also divide the space outside the region of the training set. (d) A network that presents a classification together with a confidence measure both separates the two groups and indicates 'lack of experience' for cases which are different from the training data.

achieve with a linear discriminant technique and practically impossible to create using decision trees.

### Certainty factors

Decision support systems of today, for example interpretation programs in computerized ECG recorders, present certainty factors such as 'definite', 'probable' and 'possible' (Macfarlane & Lawrie, 1989). It should be stressed that this sort of uncertainty is different from the confidence problem addressed in this study. The uncertainty arises when there is overlap between different groups, i.e., the same data can give rise to different groups. For example, if the criteria presented above, which include the limits I and II (Fig. 4a), are fulfilled, the statement 'infarct' is reported. Figure 4a shows that both two normal subjects and two patients with infarct are falsely classified. All infarcts are correctly classified when the limits I and II are replaced with slightly lower values. This increased sensitivity is balanced by a lower specificity, i.e. an increased number of normal subjects would falsely be classified as infarcts. If these criteria are fulfilled, the statement 'possible infarct' is reported. The boundaries

of the statistical (Fig. 4b) and neural network (Fig. 4c) methods can be adjusted in a similar way in order to achieve different certainty factors.

### Clinical implications

Neural networks are often developed in two-group situations, for example myocardial infarction versus normal (Hedén *et al.*, 1996), whereas the differentiation of interest in the clinical situation is infarction or non-infarction. Therefore, more comprehensive training sets, for example consisting of ECGs from normal subjects as well as patients with myocardial infarction, ventricular hypertrophy and conduction defects, are preferable. However, even a large training set is only a sample of a population and there will always be cases different from those of the training set. The confidence measure method is generally applicable and could therefore be useful in many different decision support systems.

### Conclusion

Can we rely on a black box such as a decision support system based on an artificial neural network? The quality of neural networks has been shown to be high if the quality and relevance of the training data are high (Baxt, 1995; Dybowski & Gant, 1995). The results of this study demonstrate that artificial neural networks can also indicate 'lack of experience' for cases which are not similar to any of the cases presented to the network during the training process. This method increases the possibility of artificial neural networks being accepted as reliable decision support systems in clinical practice.

### Acknowledgements

This study was supported by grants from the Swedish Medical Research Council (K97-14X-09893-06B), the Faculty of Medicine and the Faculty of Natural Sciences at Lund University, Sweden, the Swedish National Board for Industrial and Technical Development and the Swedish Natural Science Research Council.

### References

Baxt W.G. (1995) Application of artificial neural networks to clinical medicine. *Lancet*, **346**, 1135–1138.

Cross S.S., Harrison R.F. & Kennedy R.L. (1995) Introduction to neural networks. *Lancet*, **346**, 1075–1079.

Dybowski R. & Gant V. (1995) Artificial neural networks in pathology and medical laboratories. *Lancet*, **346**, 1203–1207.

Hedén B., Edenbrandt L., Haisty Jr W.K. & Pahlm O. (1994) Artificial neural networks for the electrocardiographic diagnosis of healed myocardial infarction. *Am J Cardiol*, **74**, 5–8.

Hedén B., Ohlsson M., Edenbrandt L., Rittner R., Pahlm O. & Petterson C. (1995) Artificial neural networks for recognition of electrocardiographic lead reversal. *Am J Cardiol*, **75**, 929–933.

Hedén B., Ohlsson M., Rittner R., *et al.* (1996) Agreement between artificial neural networks and experienced electrocardiographer on electrocardiographic diagnosis of healed myocardial infarction. *J Am Coll Cardiol*, **28**, 1012–1016.

Hedén B., Öhlin H., Rittner R. & Edenbrandt L. (1997) Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. *Circulation*, **96**, 1798–1802.

Johnston M.E., Langton K.B., Haynes R.B. & Mathieu A. (1994) Effects of computer-based clinical decision support systems on clinical performance and patient outcome. *Ann Intern Med*, **120**, 135–142.

Kahn C. E. (1994) Artificial intelligence in radiology: decision support systems. *RadioGraphics*, **14**, 849–861.

Macfarlane P. W., Lawrie T. D. V. (1989) *Comprehensive Electrocardiology*, Vol. 3, pp. 1540–1554. Oxford, Pergamon Press.

Pahlm O., Haisty W. K., Wagner N. B., Pope J. E. & Wagner G. S. (1991) Specificity and sensitivity of QRS criteria for diagnosis of single and multiple myocardial infarcts. *Am J Cardiol*, **68**, 1300–1304.

Peterson C., Rögnvaldsson T. & Lönnblad L. (1994) JETNET 3.0 – A versatile artificial neural network package. *Comp Phys Comm*, **81**, 185–220.

Place J. F., Truchaud A., Ozawa K., Pardue H. & Schnipelsky P. (1994) Use of artificial intelligence in analytical systems for the clinical laboratory. *Ann Biol Clin*, **52**, 729–743.

Roberts S., Tarassenko L., Pardey J. & Siegwart D. (1994) A confidence measure for artificial neural networks. In: *Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare*, 23–26 August, 1994, University of Plymouth (eds Ifeachor EC & Rosén K. G.), pp. 23–30. University of Plymouth, Plymouth.

RUMELHART D. E. & MCCLELLAND J. L. (eds) (1986) *Parallel Distributed Processing*, Vol 1 and 2, MIT Press, Cambridge, MA.

RÖGNVALDSSON T. (1994) On Langevin updating in multilayer perceptrons. *Neural Computation*, **6,** 916–926.

SCOTT R. (1993) Artificial intelligence: its use in medical diagnosis. *J Nucl Med*, **34**, 510–514.

WILLEMS J. L., ROBLES E. O., BERNARD R., *et al.* (1985) Criteria for intraventricular conduction disturbances and pre-excitation. *J Am Coll Cardiol*, **5**, 1261–1275.

WILLEMS J. L., ABREU-LIMA C., ARNAUD P., *et al.* (1991) The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N Engl J Med*, **325**, 1767–1773.

YANG T. F., DEVINE B. & MACFARLANE P. W. (1994) Use of artificial neural networks within deterministic logic for the computer ECG diagnosis of inferior myocardial infarction. *J Electrocardiol*, **27**(Suppl.), 188–193.