ELSEVIER

Artificial Intelligence in Medicine 27 (2003) 1–27

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Evaluation of inherent performance of intelligent medical decision support systems: utilising neural networks as an example

A.E. Smith*, C.D. Nugent, S.I. McClean

Medical Informatics, Faculty of Informatics, University of Ulster, Jordanstown, Newtownabbey, Antrim BT37 0QB, Northern Ireland, UK

Received 19 October 2001; received in revised form 15 May 2002; accepted 27 September 2002

Abstract

Researchers who design intelligent systems for medical decision support, are aware of the need for response to real clinical issues, in particular the need to address the specific ethical problems that the medical domain has in using black boxes. This means such intelligent systems have to be thoroughly evaluated, for acceptability. Attempts at compliance, however, are hampered by lack of guidelines. This paper addresses the issue of inherent performance evaluation, which researchers have addressed in part, but a Medline search, using neural networks as an example of intelligent systems, indicated that only about 12.5% evaluated inherent performance adequately. This paper aims to address this issue by concentrating on the possible evaluation methodology, giving a framework and specific suggestions for each type of classification problem. This should allow the developers of intelligent systems to produce evidence of a sufficiency of output performance evaluation.

Keywords: Intelligent medical decision support systems; Evaluation; Performance; Neural networks

1. Introduction

Intelligent ''decision support system'' (DSS) is a generic term used to cover many types of intelligent system that can be applied in the medical field. Clinicians see them as being black boxes and the safety critical nature of the domain requires that they be thoroughly evaluated, before they are acceptable to them; however, the weakness is that there is little in the way of laid down evaluation criteria. We focus on neural networks (NNs) as an example

* Corresponding author. Tel.: +44-778-7563-914; fax: +44-28-9036-6859.
E-mail address: ae.smith@ulst.ac.uk (A.E. Smith).

0933-3657/02/$ – see front matter
PII: S0933-3657(02)00088-X

in this paper, since these are one of the most prevalent in the domain. Also, Hart and Wyatt [44] have suggested that overcoming the issue of evaluating the outputs from NNs holds the key to evaluating other intelligent systems.

Most papers on intelligent system applications report on a single decision aid rather than comparing their system with others. Hilden and Habbema [46] argue that evaluation should be comparative and they are of the opinion that ''there is less prestige in developing evaluation tools as opposed to design tools''. This would possibly partially explain the lack of formal guidelines. The use of intelligent systems such as NNs, along with other methodologies, is not yet widespread in the domain, but it is growing rapidly as a means of handling situations that other approaches, such as rule-based systems, cannot handle. Some successful systems, e.g. PAPNET [65] for cytology screening, have been reported. Jefferson et al. [52] have pinpointed a lack of generaliseability as a problem with DSSs, but this may be dealt with in some proposed future protocols [19]. Hart and Wyatt [44] describe such systems as black boxes that have almost all the understanding, for the end-user, coming from the input data and the outputs. They state that the medico-legal responsibility may be potentially higher in intelligent techniques because of the perception that there is little input from clinicians.

The word evaluation is used inconsistently in assessing DSSs designed for clinical application. Shortliffe and Davis [94] state that generally there are three evaluation stages in the development process of any DSS; the evaluation of the output performance, the evaluation of efficacy and field evaluation, all three involving elements of subjectivist and objectivist evaluations. Subjectivist evaluation, as identified by Dowie [33], Heathfield et al. [45] and Kaplan [53], is concerned with mostly qualitative measures of organisational and human interface issues, whereas objectivist evaluation centres on the use of more quantitative measurement techniques to assess a system's effectiveness. Wyatt [111] details such an approach as utilising all the identifiable stages of the development from needs assessment through to cost-effectiveness analysis, in order to try to identify the ''truth'' at each stage. Evaluation has also been regarded by Engelrecht et al. [37] as an umbrella term with elements of verification (system functioning accurately), validation (domain knowledge accurately represented), and assessment (end-user and clinical impact) embedded in it. These definitions, however, are not generally agreed or utilised and sometimes evaluation is interpreted solely as the output performance of the system, without reference to any other aspect of the system.

The structure of this paper is as follows: we aim to define inherent performance and illustrate the need for it as a central part of evaluation of any medical system and discover how other researchers have dealt with this. To give evidence as to whether this evaluation is generally being addressed sufficiently, we carried out a structured Medline search. These aspects are preliminary to the main component of the current work, which is to suggest possible ways for inherent performance evaluation. (This is of course, within any broader framework of evaluation.) These suggestions are in response to a stated need at conferences and meetings, from colleagues engaged in the development of intelligent systems for medical application, who find it difficult to discover the details necessary to carry out a sufficiency of this inherent performance evaluation process. They cannot be expected to be expert statisticians or have direct access to this expertise, yet recognise that their systems have to meet the requirements of the domain, in order to be acceptable to clinicians.

2. Definition of inherent performance

Researchers vary in their interpretation of performance; for example, some such as Reggia [86] regard it as how efficacious the system is in the clinical setting. Others may regard it as its user friendliness, or many other functions at different stages in its development, as outlined by Rossi-Mori et al. [90] and O'Moore and Englebrecht [80], as well as overall decision accuracy. Wyatt [111] identifies two types of performance, that of in vitro and in vivo, or system performance on new test cases and the usability in field tests. Issues, such as the Hawthorne, or checklist effect, described by Campbell et al. [18], have to be taken into consideration in such a broad interpretation. We intend inherent performance here to mean all the measures which are carried out to examine how well the direct output from the system meets the "gold standard" (the correctly measured and agreed result, as recognised in the relevant domain). If no gold standard can be identified, then measurement against another system, analysing the same sample data is necessary. These measures include accuracy, precision and assessment of errors.

Performance measures as defined here can be seen as being at the core of objectivist evaluation, with all the subjectivist approaches as a shell around this. These performance measures are essentially statistical in nature. A diagram representing the objectivist evaluation of DSSs and the main elements of this with the performance core is presented in Fig. 1. Such evaluation of intelligent systems has to deal with specific issues particularly relating to: (a) the interpretation and understandability of the processes involved and whether these can be conceptualised to relate to the problem being addressed; (b) the ability of the system to deal with dynamic refinements of the knowledge based on feedback and as the environment changes; (c) the demonstration of generaliseability, that is, being

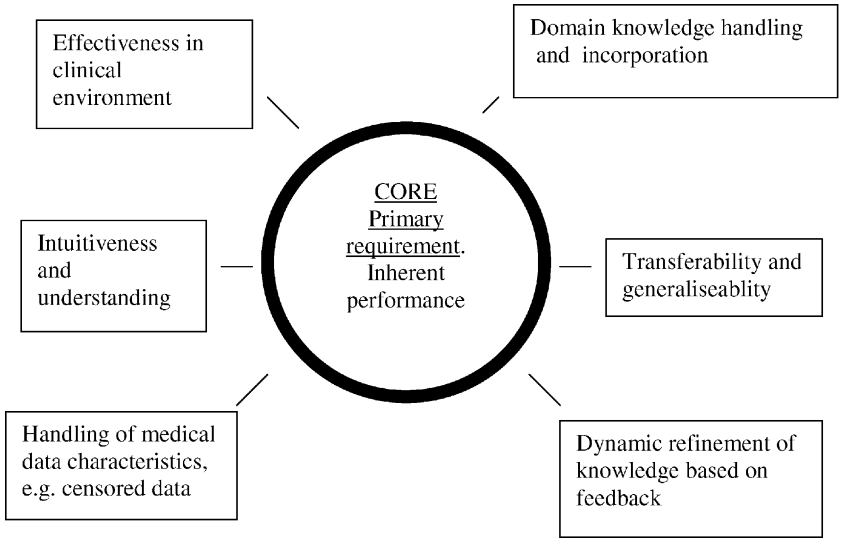

Fig. 1. Aspects of formal objectivist evaluation of decision support systems, with the core of inherent performance.

capable of being transferable to other similar environments. All this is underpinned by the core or primary requirement of the inherent performance component, which this paper deals with.

3. The criticality of inherent performance evaluation

Performance measures have been described in many papers, but these are usually related to only one type of data output. We intend here to give an overview as a basis for examining performance evaluation issues as required by the researcher for the problem in hand. Inherent performance measures, give a level of confidence that the system (or model) A is more accurate and precise than system (or model) B. Specifically, for the medical domain the advantages are that these evaluations can be shown as:

(1) Giving evidence that a real scientific approach has been applied, at least to the outputs, by overcoming the possible lack of transparency, or understandability, in the processes of intelligent systems, thus enabling clinical acceptability.
(2) More likely to offset medico-legal claims or any product liability—a growing concern for system designers, even though clinical decisions based on these should be the responsibility of the clinician.
(3) Meeting the requirements of the CE mark of the European Community.

The EU Medical Devices Directive [49,72] state that the efficacy of a medical device, including software, now has to be demonstrated, documented and clinically evaluated, and the risk to benefit ratio assessed, before a CE mark is awarded. This should include benchmarking by comparing the system with the nearest ''substantially equivalent'' approach. For many intelligent methods the use of a statistical methodology has been employed, for example, Lette et al. [63] have shown NNs to be comparable with multiple logistic regression (MLR) techniques in terms of classification accuracy.

The gold standard used as a reference for systems can be in the form of agreed measurements by experts. Any other validated and agreed objective reference, compatible with the study, can be used if appropriate. It is occasionally not possible to identify a gold standard so then direct comparisons with another approach are necessary, for example, sometimes laboratory tests give continuous output without an identified threshold indicating normal or abnormal. Gold standard sufficiency is often not achievable, but this issue is outside the scope of the paper. Suffice to say that if it is not valid, then the accuracy of classification is in doubt.

Self-testing methods for the system are not generally sufficient as this can give rise to unexpected results; for example, testing the same system on another data set can lead to unexpected problems. Also, different data gathering techniques could give dissimilar performance measures and indicate lack of transferability, a feature that Campbell et al. [19] state is a generic issue for all data mining approaches.

Researchers and developers in (a) academic institutions or (b) commercial organisations have different priorities from those in the target domain in that issues of publication and education dominate in the first case and marketing of their products dominate in the second case. The emphasis on technological methods is often incompatible with the requirements

of the end-users, that is, meeting a clearly prescribed clinical need, while researchers see critical evaluation as ''hampering their creativity''. Generally, papers in the literature on evaluation tend to deal with the broader issues of evaluation and stop short of giving specific advice on the necessary statistics to assess the outputs. Statistical and classification texts tend to be written by academics who cover the theory very well, but these texts are presented in such formal and complex mathematical language that it is difficult for non-mathematicians to understand. Statistics, however, is part of scientific methodology and can offer unassailable evidence of the superiority or at least equality of a system, so the aim is to make those accessible to the ordinary, non-mathematical researcher.

4. The overall approach

A few websites exist with the aim of giving performance evaluation guidance, such as the Statlog Project [74], where one of the aims was comparison studies of different machine learning, neural, or statistical classification algorithms. This includes the development of a software tool, ''Evaluation Assistant''. The tool is a self-testing approach, however, which only examines the one system rather than making valid comparisons, especially of measurements against the outputs of other methodologies. Other sources of information are limited in that they do not cover a full range of possibilities or are not appropriate for the medical domain.

A sufficiency of inherent performance evaluation is required and so some sort of structured framework is required to test this. Fig. 2 is the start of such an approach for the comparison of two models, showing the part evaluation of direct comparison with a gold standard, or two systems where there is no obvious gold standard, through to the full evaluation against another system via the gold standard [97]. The necessary comparisons with the gold standard, or another methodology, are almost exclusively statistical.

Confusion reigns over classification terminology, with nomenclature varying widely. Unsupervised methods, regarding the apportionment of cases by clustering, usually have no ''truth state'', i.e. no gold standard. Such a technique utilises the characteristics of features to predict a probability of a case belonging to a group. This is not likely to be presented as a medical DSS, but although no performance evaluation is possible, comparison with statistical clustering is in theory, a possibility, but this is not a true form of classification. Basically there are three types of classification. *Binary classification* is by far the most commonly encountered in the medical domain, where an example is the classification of patients into with/without the disease, based on multiple attributes, considered to be risk factors. *Multiple category classification* is the goal when there are several categories, such as several conditions, or multiple disease categories of mild, moderate and severe disease. *Continuous classification* may not always be considered to be classification but it *can* be regarded as classifying into many categories, for example, rounding up to single digits, or months estimated for survival.

Although most binary and multiple category outputs can be reduced to a percentage of classification accuracy, this causes loss of information and quoting correct classification percentages on training data only, as many papers have done, is presenting only apparent error rates. Most approaches for handling smaller data sets, which are familiar to system

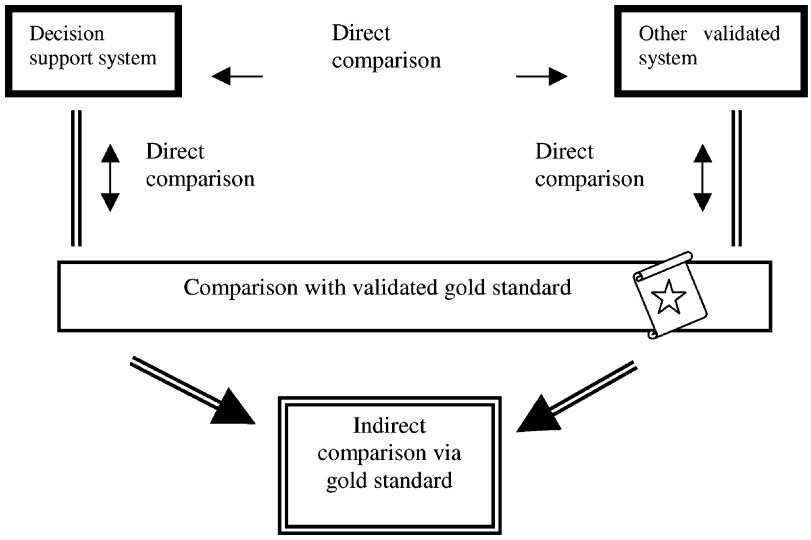

Fig. 2. A framework for evaluating the performance of a system.

developers, such as train and test, random sampling, cross-validation, and the various bootstrapping techniques, are attempts at estimating a true error rate. The accuracy of continuous classification outputs is measured in the form of mean absolute errors or the difference between predicted and actual values. Used in isolation as self-testing methods, these are best to be regarded as preliminaries to carrying out further tests on the classification performance. Score functions applied to algorithms to minimise errors or maximise their utility are not at issue here, since we are dealing with the empirical analysis of the classifier and choosing between them, rather than choosing algorithms.

This overall approach, presented in Fig. 2, indicates that the new system should be compared with other validated systems, each of which have been classified, on the same data, according to the same gold standard. Then the outputs from this classification exercise should undergo inherent empirical performance evaluation.

5. What has been utilised by other researchers?

We obtained many examples of papers on NN applications, which had been acquired because they had been referred to by other papers as having some evaluation carried out, plus many papers previously collected in the subject area. These were then reviewed for their methodological approach to evaluation.

The *binary classification* goal was prevalent in these papers, as is expected from the domain since simple disease classification is by far the most common type in medical applications. Comparisons between the outputs of NNs and many other intelligent techniques have been carried out, as well as comparisons with statistical techniques for this classification goal. Risk of disease as yes/no comparisons of logistic regression and

NNs are now fairly common. Indeed binary logistic regression is now the main corresponding statistical method of choice for comparison with NNs, in predicting those cases likely to develop a condition, given multiple risk attributes. This approach has been utilised by Deligdisch et al. [28], Biagiotti et al. [12] and Jefferson et al. [51]. Also, Lette et al. [63] have used this comparison approach to predict risk of complications after disease has occurred.

A large body of work has addressed the prediction of prognosis for mortality. Ebell [35] has compared the predictive value of NNs with a progress after resuscitation (PAR) score for assessing possible cardiac arrest patients (modified from pre-arrest morbidity (PAM) score). TRISS, a trauma survival logistic regression model has been used by Hunter et al. [48], for the prediction of prognosis for accident patients. Other prognostic models for comparison are acute physiological and chronic health evaluation (APACHE) models I–III and simplified acute physiology scores (SAPS) and mortality probability models (MPM II), as suggested by Lemeshow et al. [61]. Most of these are incorporated in logistic regression models for application in the specific domain. Backward stepwise logistic regression approaches have been used frequently, e.g. by Dybowski et al. [34], for comparison with NNs for outcomes of critically ill patients with a spectrum of diseases.

Linear discriminant analysis has been applied by Anderer et al. [5] to compare performance of NNs with respect to diagnosing demented patients on the basis of EEG activity. This has also been used by Cohen et al. [23] to compare classification of autism by NNs. Another example of a binary classification problem, utilising NNs and statistics has been that of survival analysis. Outputs from NNs have been compared with that from Cox's regression by Mariani et al. [70] who took a model of risk factors and used Cox's regression to predict single hazard ratios for survival, for each case. Factor analysis has also been used for discrimination comparisons with NNs with respect to identifying those at risk of diabetes or not, such as that carried out by Tafeit et al. [101].

For evaluation of the performances of binary classification comparisons, some of these researchers have used statistical tests on top of self-testing model development. Many of these have been evaluated with a quoted percentage of correct classification, whilst many have summarised results into 2×2 contingency tables or confusion matrices, and often presented a χ^2-statistic for the differences in proportions correctly/incorrectly diagnosed. Measures of specificity and sensitivity have had widespread use amongst researchers, e.g. Lette et al. [63] and Marble and Healy [68], and some, such as Ronco [89] have used positive predictive value (PPV) and negative predictive value (NPV). Fuller ROCs are increasingly being applied by researchers such as Anderer et al. [5], whilst Mariani et al. [70] and Doig et al. [32] have used area under the curve (AUC) measures for direct comparisons. Tafeit et al. [101] have employed a two independent samples *t*-test to compare NNs with factor analysis output. *Multiple classification* output comparisons have been carried out to compare multiple category goals, and hence the outputs from several nodes in NNs, with statistical techniques. There are less comparisons in this type of classification, but Deligdisch et al. [28] have compared NNs with a Kruskall–Wallis test and Dunn procedure for the multiple classification of ovarian dysplasia. Ohno-Machado [81] has compared the result of NNs with Cox's regression to predict the hazard ratios for individual cases, for four times categories of output, in the risk assessment for coronary

disease. Arana et al. [7] have used binary logistic regression and NN comparisons in the classification of multiple types of malignancy in bone lesions, treating each category as a binary outcome.

Evaluations of these have included χ^2-tests for k groups and proportions of successes in each group, as utilised by Ohno-Machado [81], Wright and Gough [110] and also de Laurentis and Ravdin [27]. Others such as Lapuerta et al. [60] have employed Z-statistics for two paired proportions and McNemar's test for correlated proportions. Many researchers have treated each category as an individual binary node and thus sensitivity and specificity for each group has been utilised by Azuaje et al. [9]. Fuller ROCs have also been utilised and some, such as Arana et al. [7], have given AUCs, plus standard deviations.

Continuous classification for non-parametric continuous output has been considered by Meting and Coenegracht [73] in which they compared NNs with quadratic regression techniques—a multiple non-linear (polynomial) approach to modelling chromatography optimisation. Rehman et al. [87] also compared NNs and non-linear regression modelling with a goal of non-parametric outputs that predict depth of anaesthesia and produced graphical output on this. Most continuous classifications have been in survival or time-to-event analysis. For example, Anand et al. [4] compared several NN approaches with Cox's regression, where there were censored cases in a colorectal cancer cohort of patients. Kaplan–Meier graphs and NNs were the subject of Biganzoli et al. [13], who produced graphs for both techniques for survival, again with censored data.

These continuous classifications have been evaluated with Wilcoxon's sum rank test by Anand et al. [4] and also Nugent et al. [78]. Most others have not carried out specific empirical tests, e.g. Biganzoli et al. [13] apart from giving graphical output and mean squared errors.

6. To what extent is inherent performance evaluation being carried out?

We decided to get an unbiased assessment of the degree of evaluation being carried out in the literature. To this end we utilised PUBMED to access Medline to carry out a structured search of the literature published on NNs in the medical domain.

6.1. The Medline search

The criteria used for searching for papers were—MESH term ''neural network'', title word: limitations—English language; human; publication dates 1 January 1990 to 31 December 2000. This produced a total of 1140 papers and a systematic random sampling of these resulted in 56 apparently suitable papers, which were retrieved in full from libraries or electronic sources. Eight of these were found to be not artificial intelligence papers, so the overall result was 48 papers extracted [6,8,10,12,14–17,21–23,28,31,39,40,42,50,51,55,57–59,62,64,66,67,69,71,76,82–84,88,91–93,95,99–104,107–110,112]. We would point out that these articles do not include preliminary reports or conference papers, but only these that are now in the general literature and have been peer reviewed for publication.

6.2. The proforma

The papers were examined, the reference details (a) and (b) noted, and the answers recorded in a proforma (Appendix A), which we designed with the view of answering the following questions.

(c) What was the intention of the authors? Was it intended that this reported research could go on to be applied in the medical domain or not, or is it purely of academic interest? Statements such as "ultimate aim of introducing it into routine use" or "powerful methodology for aiding the expert" or a clear application as a model for analysing medical data were interpreted as having an intention to promote the application of the method. The intention to give details of methodology that could be developed further was interpreted by certain statements in the paper. These included "a possible future application", a "prototype system", "should be developed further" or that "no evaluation is necessary", since the approach is purely theoretical at the current stage.
(d) Was there an identifiable gold standard? Gold standards raise the issue of sufficiency (cf. comments in Section 4). Clustering, of course, does not usually have a gold standard, unless a supervised comparison has been made.
(e) What broad type of evaluation was carried out? (All the comments in Section 4 are relevant here.) Comparison with a gold standard or with another methodology? Comparison with another methodology via the gold standard? The validity of all these comparisons? For example, sometimes the comparisons with another system were invalid such as comparing non-linear methods (NNs) with linear methods. Were the methodologies of the comparisons sufficient? (Comparison methods themselves range from simple correlation coefficients through to full appropriate analysis giving a direct quantitative and valid comparison of the inherent performance between the two systems.) An attempt at scoring is presented from: (0) No evaluation. (1) Self-testing—train and test, cross-validation techniques; comparison with gold standard that was thought to be insufficient, e.g. percentage classification error or root mean square only; or invalid comparison with a dissimilar methodology, such as comparing linear with non-linear methods. (2) Against a validated gold standard, with some rigour and statistical analysis. (3) Comparison with another system, but either not quite the same data or scenario or population or gold standard. (4) Against another methodology via a validated gold standard, but insufficient, e.g. MAEs quoted. (5) As (4), but specificity and sensitivity quoted. (6) Fuller evaluation, e.g. areas under ROC curves, full pairwise analysis.
(f) Finally, were any of the above related to the type of classification problem, i.e. no classification (clustering) or the three categories defined in Section 4?

Results of applying the proforma to the 48 papers are presented in Tables 1–3. Table 1 presents the frequency of each category of evaluation as defined above. This indicates that 12.5% of papers gave what might be a sufficiency of evaluation ((5) or (6)). Most papers (79.2%) did not describe comparison of their system or model against another system ((0)–(3)). The rest (8.3%; (4)) attempted to compare their system with another, but used inappropriate statistics.

Table 2 presents an analysis of the results of the structured Medline search. The $r \times c$ cross-tabulation indicates that the degree of evaluation in the paper is associated with the

Table 1
Table of frequencies of evaluation in the Medline search

Degree of evaluation		Number	Percentage
0	None	4	8.3
1	Train and test, cross-validation (self-testing) *or* against gold standard, poor *or* against invalid other methodology	27	56.3
2	Against gold standard, good	4	8.3
3	Against other system, poor, *or* not via same gold standard	4	8.3
4	Two systems, via gold standard, but poor	3	6.3
5	Two systems, via gold standard, just sufficient	4	8.3
6	Two systems, via gold standard, full, appropriate	2	4.2

Table 2
Analysis of evaluation in the Medline search

Degree of evaluation with	Pearson χ^2	d.f.	Sig.	Exact Sig.
Intent	15.33	6	0.018	0.022
Gold standard	41.98	6	<0.001	<0.001
Year	55.91	54	0.403	0.403
Class type	39.10	24	0.027	0.066

d.f., degrees of freedom; sig., two-sided significance; Exact Sig., an extended Fisher's exact test to account for low counts in some cells.

intention of the authors (intent) to apply the model or not ($P = 0.022$). The degree of evaluation is, unsurprisingly, associated with the presence of an identifiable gold standard ($P < 0.001$). The year of publication has no effect on the degree of evaluation found ($P = 0.430$), indicating that the degree of evaluation has not increased over the sample period, i.e. 1990–2000. The classification type category was marginal with respect to significance feature ($P = 0.066$) and this has been expanded to present the results indicating the type of classification and the degree of evaluation in Table 3.

Table 3
Cross-tabulations—classification type with degree of evaluation Medline search

Degree of evaluation	Classification type			
	a	b	c	d
0			1	
1		13	12	1
2		2	1	1
3	1	1	2	
4		2	1	
5		4		
6		2		
Total	1	24	17	2

a, no class; b, binary; c, multiple; d, continuous. The paper in (a) was a clustering techniques paper. Three papers were in a category of ''other'', e.g. feature selection techniques.

Table 3 indicates that just over half (13/24) of the most common type of classification—binary (b)—were of degree 1 evaluation. Otherwise 6/24 fell into the categories of sufficient evaluation (degree 5 or 6) and the rest were divided between the other evaluation categories. Of the second most common classification type—multiple (c) 12/17 were of grade 1 evaluation, none in the full evaluation categories of 5 or 6 and the rest divided between the other categories. Only one paper dealt with no classification and only three with continuous data outputs, which were evenly distributed in the evaluation categories. This makes a statement that the modal evaluation for both binary and multiple category classification is degree 1, i.e. poor evaluation.

6.3. Discussion of Medline results

The aims of this part of the study were not to give a precise breakdown of the paucity of inherent performance evaluation, and criticise papers in the literature. We only wished to indicate that there might be an overall lack of stringency in the approach to inherent performance evaluation, especially given that these systems are being suggested as suitable for application in the medical field. This conclusion has been justified by the study.

7. A methodological overview of suggested possibilities for inherent performance evaluation

We have used the evidence from the literature review of Section 5 along with our own expertise to try to construct a methodology overview for this topic. Full formal performance evaluation, as defined in Sections 4 and 6, is all about valid comparison with another equivalent methodology, which can be all types of artificial intelligence techniques, plus Bayesian networks and statistical approaches, or a hybrid of these. (Bayesian methods are considered to be systems on their own and generally require specific expertise in the field; however, simple Bayes' theorem is integral to many so called ''Frequentist'' tests.)

When testing for performance, it is better to use the raw data output for each case in the set, wherever possible, rather than producing an overall classification in the form of a contingency table or confusion matrix. Mistakes can be made when the data is summarised and information lost, plus no details are given on the shape of the distribution of the output. Comparisons of classifications are best carried out on paired output, but it is recognised this is not always possible. Primarily the focus should be on what alternatives are being considered in the analysis and goodness of fit tests. Because this is inherent performance evaluation, we are not concerned with costs and benefits or opportunity costs or clinical impact, as these are evaluations at a different stage of the development of a system. Nor are we concerned with building the model here, or the score functions to minimise the errors of the algorithm, although comparisons of performance may give feedback for this. When testing at the level of train and test sets, whether leave-one-out, 10-fold cross-validation, two-third and one-third, etc. to give mean squared errors from the gold standard, these divisions of the data can be compared directly with the same divisions of the statistical or other system model.

A complete approach to inherent performance does not exist. We will attempt to give researchers and developers of systems some indications of the possible valid comparisons

and tests that are appropriate to each classification goal, using NNs as an example. The aim is to give information to those who have little idea as to how to go about evaluating their system, rather than the expert. This will attempt to cut through the maze of information to give specific suggestions, which may be applicable to the problem that the researcher is trying to overcome. All testing can be applied at the cross-validation stage or to the whole model stage, providing it is between two classifiers, under the same conditions. The distinction is not hard and fast, it is essential to keep appropriateness in mind, rather than proscribed constraints. This said, mean squared errors, Brier scores, etc. are often quoted by researchers as being evidence of performance, but these are insufficient on their own as performance measures. Also inequalities stating that >1 or 2 standard errors (S.E.) are significant, whereas a conventional *t*-test would give specific probabilities and thus more opportunities for direct comparisons.

We suggest that questions on the nature of the output from the system should be addressed in a decision tree type of approach. Firstly, is the output unsupervised or supervised in the form of a binary, a multiple category or a continuous output? (Unsupervised clustering methods cannot be evaluated properly as such, in the absence of any gold standard, but informal comparisons with statistical clustering techniques are possible.) However, if clustering is of a supervised nature as, for example, with some k nearest neighbour (knn) approaches, where a gold standard exists, then evaluation is possible and the output can be treated as one of the other three types of classification and tested as such. (This has been carried out by Anand et al. [4].) Deciding the form of the classification can be difficult for the researcher as some think that all classification can be considered to be binary, depending on the point of view. Nugent et al. [79] have shown that it is possible for classification problems to be segregated into a number of bi-dimensional classification problems. This issue relates to the goal of the classification and the way the output data is to be presented. Although this may appear initially confusing, the researchers will understand the nature of their own data and that being produced by the output of the system and be able to interpret accordingly. Within these categories, is the output in the form of a series of raw data points or presented as contingency table? Is the output data spread parametric or non-parametric, that is, is there a recognisable distribution form or not?

We will take each classification in turn and suggest suitable comparisons and tests for the outputs. Since space is limited in one paper, these will necessarily be very brief and only an indication of what is possible. For each classification goal, possible comparison methodologies are given and then tests of performance suitable for this type of goal. These are discussed in the text here, but the intrinsic nature of the subject means that it does not lend itself to readability, so for a fuller interpretation of these, we suggest the reader refers to suggested texts for further explanation on each topic as required. The methodologies and tests are in *italicised* text, to highlight them and some knowledge of standard statistical shorthand is assumed.

7.1. Binary classification

This is by far the most commonly used classification goal in medical applications, typically required for the discrimination between disease or not disease categories, or those

at risk or not. For example, single nodes in output layers of NNs frequently are binary in nature with a threshold identifying a yes/no division in the predictor attribute.

7.1.1. Binary classification—comparisons

Researchers have compared NNs with *linear discriminant analysis* (cf. Section 5), but this is more appropriately used for grouping into three or more categories of dependent attributes and is relatively complex for non-statisticians to apply. It also lacks flexibility and generally requires normality of distribution of the predictor attributes, although a little flexibility with some binary attributes is possible. When there are just two groups for classification, Vach et al. [105] state that similar results to discriminant analysis can be achieved using binary *multiple logistic regression* (*MLR*) and NNs. A clearly presented example of a comparison between these two methodologies is contained in the paper by Arana et al. [7]. Ely et al. [36] state that MLR gives greater flexibility, readily incorporating models with quadratic terms and multiplicative interaction terms, whilst having the benefit of being more easily understood. Both NNs and MLR are similar iterative processes and MLR can be regarded as using weighted scoring of all the inputs, in an analogy to that of NNs. The output of MLR is constrained between 0 and 1, by a logistic transformation and this can be interpreted as the probability of membership of that class, since this basically follows one of the same forms as the NN output—a logistic regression function:

$$\text{estimated probability} = \frac{1}{1 + \mathrm{e}^{-z}} \tag{1}$$

where e is the constant 2.718..., from $\log_{\mathrm{e}} x$. The z is the linear combination of the covariate values: $z = B_0 + B_1X_1 + B_2X_2 + \cdots + B_pX_p$.

Traditionally, statistical models are reduced to give a parsimonious model, but we believe that all attributes should remain in the logistic regression model as far as possible for comparison with NNs. If these have to be reduced, then we advise cutting down the model for both techniques by using, e.g. χ^2-tests, explained by Delucci [29] for each factor individually for the two classes as identified by the gold standard. Stepwise logistic regression is also used frequently for feature selection. A reasonably safe approach is to take each attribute that has $P > 0.1$ out of the model, but leave the rest, even if not significant, in order to retain as far as possible those that may contribute to interactions. It is necessary to compare like with like, that is, duplicate the model for all the techniques otherwise the comparison may be invalid. For the same reason we suggest that, for the final model, researchers enter all the attributes together in the MLR model, since this is the same as the input layer to NNs. Other researcher may have used both forward and backward stepwise regression procedures, but Derksen and Keselman [30] demonstrate that there can be problems with stepwise regression for comparison, for example, when multiple testing is required.

Simple survival analysis where all the cases are uncensored, where this means that there is a record of survival or the event is known, can be compared with MLR. If there are censored cases where the event is unknown, or lies outside the time frame of the study, then this can be compared with *Cox's regression* [25] for covariates. This gives prediction of binary hazard ratios for survival, or event-or-not, but all the previous comments on similarity of modelling between the two systems apply here. Log-rank tests for simple

models with few covariates can also be employed. A full explanation of this and other survival modelling techniques are given in Collett's textbook [24] on the subject.

All the risk scoring and models, plus mortality prediction (prognosis, without censoring), mentioned in Section 5, which are mostly derived from MLR modelling, can be used for comparison with NNs. These examples have the advantage of being established in the domain, but exactly the same attributes and weights of these models need to be presented to the input layer.

As a simple approach, it is possible to compare binary output with that of Z-values, when the data is transformed to achieve a normal distribution and $z = (x_i - x_{\text{mean}})$/standard deviation, as Anderer et al. [5] have done. For example, are the proportions produced by the classifier different from the proportions that would be produced by the appropriate Z-values on the same data?

7.1.2. Binary classification—testing of performance

The evaluation of binary classification has received more attention than any other type, and many multiple node outputs are treated as a series of binary outputs for evaluation purposes. The researcher has to decide which output classification is appropriate to the questions they are trying to answer. For example, cross-fold validation results can have simple *t-tests* applied ([20], p. 132–4) for the differences in mean absolute errors, using:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{S.E.}(\bar{X}_1 - \bar{X}_2)} \tag{2}$$

the hypothesis of no difference in the means. These are usually in the form of testing of paired samples, where the data has the same cases being compared, the paired t-test. They can also, however, be in the form of independent samples where these are not exactly the same and adjustments have to be made to the standard error to account for this. The t-test assumes that the data is approximately normally distributed about the mean error and is useful for small samples (<30) in that it depends on the size of the sample and thus the degrees of freedom for the resultant probability [2]. A note about this would be to avoid just stating the inequality, often quoted in the intelligent system literature that if the value of t is ≥ 2, then it is significant, without giving an exact probability value. If the output data to be tested is large, i.e. >30, then the difference in proportions between two binary outputs can be tested with Z-tests [26] where this is asymptotically the same as the t-statistic:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{(\text{S.E.})_{\text{difference}}} \tag{3}$$

This utilises the standard error and gives a test statistic and associated probability.

If the data is presented as a 2×2 contingency table, then researchers should decide if the samples require a paired proportions test. Perhaps the most appropriate test here is *McNemar's test* on a 2×2 contingency table of comparisons of ratios ([3], p. 258–9) if the output data to be tested is in this reduced form. If the comparisons, between methodologies, are not on exactly the same data then they will be independent and unpaired and require the use χ^2*-test for* 2×2 *tables*, which is testing for non-paired proportions:

$$\chi^2 = \sum \frac{|O - E|^2}{E} \tag{4}$$

Then employment of *Yates continuity correction*, for bias produced by small samples:

$$\chi^2 = \sum \frac{(|O - E| - 0.5)^2}{E} \tag{5}$$

and *Fisher's exact test* if any cell in the table has frequency <5, where the probability of the other tables giving the same marginal totals are summed to give more evidence of association. McNemar's and χ^2-tests are covered in Altman's textbook ([3], p. 250–9).

In supervised learning, where a possible threshold can be identified whereby the output measure can be deemed to be a correct classification or not, *ROC analysis* in all its forms is appropriate, where there are positive (+ve) and negative (−ve) values, that indicate the outcomes with respect to the classification. *Sensitivity and specificity*, which are part of this, are regularly quoted in the literature as evidence of performance evaluation, along with the positive predictive value (PPV) and negative predictive value (NPV) [43] with the following definitions.

- *Sensitivity*: proportion of +ves which are identified by the system.
- *Specificity*: proportion of −ves which are identified by the system.
- *PPV*: proportion of true +ves which are identified by the system.
- *NPV*: proportion of true −ves which are identified by the system.

However, some researchers may not be aware of the shortcomings and constraints of these, such as the prevalence of disease in the population or the prior probabilities of the condition or state, which will affect the classification threshold. This involves Bayes' theorem calculations for PPV and NPV based on these prior probabilities to give posterior probabilities ([106], p. 244–6).

A likelihood ratio (LR) can be calculated ([3], p. 416–7) which can perhaps simplify matters, where LR = sensitivity/(1 − specificity). This can be combined with what can be termed as pre-test odds, where this is the prevalence/(1 − prevalence), to give the post-test odds, which are effectively the change in certainty of the classification from that pre-existing to that predicted by the model:

$$\text{conditional probability} = \text{pre-test probability (odds)} \times \text{LR} \tag{6}$$

These conditional probabilities (or proportions) can then be compared directly, using a χ^2-test.

If exactly the same model and exactly the same population are being used, however, then this should not influence the test results of a direct comparison between techniques. The way to ensure against this is to carry out full ROC analysis; which provides a rigorous assessment of a test's diagnostic accuracy and is independent of the prior probabilities. ROC curves display the ability of a classifier to discriminate between binary groups without having to specify a threshold level, by graphically displaying the sensitivity versus the *false positive rate* (1 − specificity) for all possible threshold values produced by the classifier. In this way, they give summaries of performance over a range of possible decision cut-offs and are therefore insensitive to the selection of these, thus avoiding the erroneous results that can occur with stipulated thresholds.

For direct comparisons of ROC curves, the *area under the curve* (*AUC*) is increasingly being used [11] and these can be considered to be a measure of the diagnostic power of a

test, independent of thresholds chosen by the researcher [113]. The significance of the differences between areas are more complex to calculate but *Pearson's correlation coefficient* which measures the degree of association (−1 to +1) between the values of the two variables/attributes, can be used. Otherwise *Kendall's tau*, a rank correlation coefficient when there are small numbers of categories, can be employed. Both are suggested by Hanley and McNeil [43]. These methods give significances at a chosen level (usually $\alpha = 0.05$) by:

$$Z = \frac{A_1 - A_2}{\sqrt{(\text{S.E.})_1^2 + (\text{S.E.})_2^2 - 2r(\text{S.E.})_1(\text{S.E.})_2}} \tag{7}$$

where r is one of the above correlation coefficients and Z the normal score to be converted to the probability. Many statistical packages now give the area under the ROC curve calculations, so quoting these has become simplified.

7.2. Multiple category classifications

The ideas of Section 7.1.2 for testing of binary outputs can generalise readily to the multiple category classification case. Often it is useful to draw up a contingency table or a confusion matrix, a cross-classification of predicted versus the observed class. When there are, for example, multiple nodes in the output layer of NNs then comparisons can be carried out with statistical techniques that classify a dependent attribute into several categories. Keeping a clear focus on the emphasis of the analysis would answer the following questions. Is each node to be considered as a separate binary classification or separate continuous classification? Are all the outputs related as multiple categories and are inter-category comparisons required?

7.2.1. Comparisons of multiple category classifications

Discriminant function analysis [5] can be used to predict group membership, but this is usually possible only when predictor attributes are normally distributed or nominal, but not of equal variances and where the dependent attribute is a nominal measure. This approach cannot handle a dependent attribute with a large number of categories. Researchers should remember to use the ''priors'' option if there is a prior knowledge of the proportions in each category from the data set being examined and if this has been incorporated into the application. Other researchers see this analysis technique as being difficult for non-statisticians to handle and interpret.

Perhaps the most appropriate comparison with systems such as NNs with multiple nodal output, is that of *multinomial logistic regression* [98], which can handle a dependent (predictor) attribute with few categories, but otherwise is similar to ordinary multiple (binary) logistic regression where parameter estimation is performed through an iterative maximum likelihood algorithm. This procedure is little utilised in intelligent systems research as yet. For survival analysis problems where all the cases are uncensored in the form of multiple category classifications of time, e.g. 0–6 months, 1–2 years, etc. are required, it is possible to compare the outcomes with multinomial logistic regression. Otherwise, if there a censored cases, the outputs could be compared with hazard ratios for

survival with *Cox's regression* [25], using these same categorisations as the system output for separate testing. The censored cases can be inserted into the final category, as other researchers have done [82].

7.2.2. Multiple categorical classification—testing for performance

Generally, χ^2*-tests for* $2 \times k$ *categories*, as used by de Laurentis and Ravdin [27], in:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{8}$$

the form of contingency table tests are useful here, but we would suggest the employment of a χ^2*-test for trend* ([3], p. 261–5) if there are ordered categories, which will give a more powerful test because it gives a test statistic in the same way, but on one degree of freedom rather than $k - 1$ degrees of freedom on the χ^2-distribution. Testing with *two-way analysis of variance* (*ANOVA*) ([41], p. 191–4) can be carried out if the outputs are individually approximately normally distributed (a generalisation of the *t*-test). Also *Friedman's analysis* ([3], p. 334–6), which is the non-parametric equivalent of this can be applied if visualisation of the output indicates that the assumption of normality of distribution is invalid. Sometimes inter-rater agreements can be applied to categorical outputs; *Cohen's kappa* a measure of ''chance-corrected proportional agreement'' ([41], p. 187–8) will give values between -1 and 1. This measures the agreement between estimates for the same categories and ranges from worse than chance agreement (-1), through no agreement greater than chance (0), to perfect agreement (1):

$$\text{Cohen's kappa } (\kappa) = \frac{O_{\text{Ag}} - E_{\text{Ag}}}{1 - E_{\text{Ag}}} \tag{9}$$

If the outputs are to be considered to be independent individual binary outputs, then all the techniques employed under Section 7.1.2 can be utilised. These include *Z*-statistics, or McNemar's test, or χ^2-test for 2×2 tables, or sensitivity and specificity for each category, or ROCs and AUCs for each category (cf. Section 7.1.2).

Alternatively, if the output categories are considered to be individually of a continuous output nature, then all the tests under continuous outputs, which follow, could be applied.

7.3. Continuous classification

Continuous classification can be in the form of, e.g. laboratory testing or any other scale that does not have a specific threshold that identifies normal/abnormal or other binary class results. If a threshold can be identified then all the binary comparisons and testing above can be applied. The most common form of continuous classification used by system developers has been in the estimation of survival or time-to-event data.

An elementary preliminary for continuous classification output is to consider whether this is in the form of a parametric or non-parametric distribution. Testing for the type of distribution should be carried out before further analysis. Tests include those for normality, e.g. using *Shapiro–Wilks test* ([3], p. 139–42), but check for outliers as this test is very sensitive to these. Other parametric distribution fits to the data, e.g. using *curve fitting*

approaches [1], can quickly reveal other possible distributions found in continuous data in the medical domain, such as the Weibull distribution or negative binomial distribution. Linearity of the data is also detected by curve fitting and measure of R^2, the line fit.

7.3.1. Continuous classification—comparisons

Comparisons of this type of system output are dependent on the shape of the distribution. If the dependent attribute can be shown to be a linear value that represents an additive association of the predictor attributes, which themselves are normally distributed, then *multiple linear regression* is appropriate ([20], p. 93–6):

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \tag{10}$$

where y is the linear combination of covariates (βx) and α is a constant.

If non-linear, then the equivalent of this, a multiple non-linear (polynomial) form such as a *quadratic regression* model ([3], p. 310) can be used:

$$Y = a + bX + cX^2 \tag{11}$$

When investigating survival or time-to-event analysis and where there are no censored cases, comparison with *multiple linear regression* is possible if the assumptions for this hold. When there are right or left censored cases, comparison of survival estimates can be made with *Kaplan–Meier* [54] graphs and estimates for time intervals, provided there are few attributes being considered, using:

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} \tag{12}$$

where n_j is the number still surviving at the jth time start and d_j is the number who have the event in the jth time interval. $\hat{S}(t)$ is the probability of surviving until after time t. If covariates are available and the assumptions of constant proportional hazards with time can hold, then *Cox's regression* ([24], p. 54–5) can be utilised in the form of survival estimates rather than hazards to survival:

$$\hat{S}(t) = [S_0(t)]^p \quad \text{where } p = \mathrm{e}^{\sum B_n X_n} \tag{13}$$

where X_n are the covariates. Direct comparisons can now be made using the technique of Smith and Anand [96] to give actual time estimates for each case, for direct comparison with intelligent systems which give an estimate of time-to-event. Where the assumptions required for Cox's regression are seriously eroded, then Weibull analysis can be used if the system output is in the correct distributional form for this ([24], p. 110–3).

7.3.2. Continuous classification—testing for performance

Generally continuous classification does not have +ve or −ve categories, that is, there is no decision threshold and so ROC curves and all the tests under this are inappropriate. If, however, it is possible that a threshold could be used for classification into binary categories, then all the testing under Section 7.1.2 is applicable. If the system output is normally distributed, then the obvious approach is to carry out *paired t-tests* (Eq. (2)) ([20], p. 132–4) using $n - 1$ degrees of freedom, rather than just comparing the mean absolute

errors. Also, *independent sample t-testing* (Eq. (2)) can be utilised for non-paired data outputs, using $n_1 + n_2 - 2$ degrees of freedom, which gives a less powerful test than the paired samples. Researchers should give confidence intervals and standard errors where appropriate. With linear outputs, if the degree of association between two outcomes is required, then *Pearson's correlation coefficient* [56] can be calculated from the original observations, but because of restrictions on the validity with this, it should be considered to be a preliminary analysis only. Non-parametric equivalent coefficients are available in the form of the various rank correlation coefficients, such as *Spearman's rho* ([20], p. 82), and *Kendall's tau* [56]. These have the advantage that confidence intervals can be constructed to give more information rather than just a coefficient and a probability.

Non-parametric testing of paired outputs mainly uses *Wilcoxon's rank sum test*:

$$Z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{14}$$

where n is the number of pairs that are ranked according to the absolute values of the differences between the two attributes. This has been used by Anand et al. [4] to directly compare the outputs from NNs, knn models and regression tree induction with each other and with Cox's regression, for survival analysis, where the output was not normally distributed. Care is required in interpretation of this ranking test, however, because this does not take account of the size of the differences—it only ranks them by magnitude comparisons, so it may give different results to, for example, the mean absolute errors. If the output is unpaired, the *Mann–Whitney U-test* ([20], p. 141–2) should be applied, it being the non-paired equivalent of the Wilcoxon's test:

$$Z = \frac{T - n_1(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \tag{15}$$

Kaplan–Meier and Weibull graphs can be compared by using the *log-rank test* ([24], p. 40–3), based on a χ^2-distribution, for individual graphical output, where A and B are the two graphs being compared:

$$\chi^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} \tag{16}$$

Note: Parametric tests are more powerful and should therefore be used in preference to non-parametric tests when the assumptions (e.g. normality) hold. Note that confidence intervals and standard errors should be given wherever possible. Also note that if several separate *t*-tests are carried out then researchers should employ *Bonferroni's modification*, where the probability obtained from each test should be multiplied by the number of separate tests ([3], p. 211), however, care in interpretation is required [85].

7.4. In summary

As stated previously, this paper is intended as an indication of what is possible, but may save the researcher much searching, by suggesting suitable methodologies for comparisons of classifiers. The summarised form of this is presented in Table 4.

Table 4
Possible tests and comparisons according to classification goal

Type of classification goal	Possible comparisons	Possible tests
Binary	*Z*-statistics Linear discriminant analysis Multiple logistic regression	Independent sample *t*-tests Paired *t*-tests *Z*-tests McNemar's test χ^2-test for 2 × 2 tables Sensitivity, specificity Positive and negative predictive values True, false positive values Area under the ROC curve (with Pearson's correlation coefficient or Kendall's tau)
Multiple category	Discriminant function analysis Multinomial logistic regression	χ^2-test for 2 × *k* categories χ^2-test for trend Two-way ANOVA Friedman's analysis Cohen's kappa
Continuous	Multiple linear regression Quadratic regression For survival or time-to-event data Kaplan–Meier graphs Cox's regression, Smith adaptation Weibull analysis	Independent sample *t*-tests Paired *t*-tests Pearson's correlation coefficient Spearman's rho Kendal's tau Wilcoxon's rank sum test Mann–Whitney *U*-test Log-rank test

We have given indications of suitable tests that should be further investigated, if the researcher is unfamiliar with them, for constraints and limitations. Suitable sources giving some of the mathematics involved can be found in textbooks such as Altman [3], Campbell and Machin [20], or other sources referred to in the text. Survival analysis types can be examined in Collett [24]. All the conditions and constraints should be taken into account along with knowledge of the domain, the data being used, the problems present and the questions being posed.

8. Discussion

Many evaluation papers for intelligent systems designed for medical application have been published, mostly of a subjectivist nature, e.g. [47,75,77,80]. The reader could also look at Friedman and Wyatt's "Evaluation methods in medical informatics" [41] to get the global evaluation perspectives, both subjectivist and objectivist, and the issues involved. Also, Ergemont-Petersen et al. [38] have specifically examined the quality and performance of NN classifiers and this is pertinent to this subject. Objectivist studies have

indicated the possible quantitative ways to evaluate systems, but have, however, not generally given structured, comprehensive information about how to carry out adequate inherent performance evaluation, according to the classification goals of the outputs, as this paper has tried to do. This seems to be largely left to those who are interested in mathematical classification theory and are published in a form not easily understood by the non-experts in the field.

There is a general recognition that new technologies such as NNs should be medical problem driven with output that is both appropriate and understandable to the end-user. Clinicians are wary of new inroads into their domain and so it is advisable that systems should be compared with statistical and other current acceptable approaches utilised in the domain. The gold standard relevant to the DSS goal should be sufficient and, as stated earlier, not just the diagnosis of one clinician.

Randomised controlled trials (RCTs) are used routinely in medical research, where they perform well in the original concept of direct comparison of treatment regimes. They are now being regarded by some, however, as generally unsuitable for the evaluation of intelligent systems, since they take little account of the barriers to the introduction of new technologies and are also limited in their range of coverage [45]. Others would disagree and state that they could be applied to more complex interventions in an iterative approach, gathering all the results from all the elements together [19].

There are no ideal answers, but this is not to say that an attempt cannot be made to test the inherent performance of systems empirically. Our intention is to give details of the specific tests appropriate to the type of data output, along with examples and constraints. Rossi-Mori [90] is of the opinion that these measures should essentially treat the system processes as a black box, as the methodological processes are hidden at this particular stage of performance.

We recognise that the imposition of too judgmental a form of assessment criteria, while the system or model is being finalised, would mean that many systems and ideas would not reach all the intended recipients who might benefit from it, or develop it further. If it is intended that the system or model be applied in the domain, however, then adequate formal performance evaluation is necessary. Not carrying out this core evaluation avoids answering—"does the system do what it claims?" and "is it more accurate than current methods?" Such questioning is essential for giving evidence that a real, scientific process has been applied to meet the safety—critical requirements of medical systems. All this is part of an overall evaluation of these systems, with all the subjectivist and objectivist components mentioned in Section 1.

9. Conclusions and future work

What can be seen to be new in this paper is an attempt to organise a practicable approach to performance evaluation of DSSs in a manner that designers of system themselves can use, to carry out some adequate evaluation. This is a sufficiency of methodology rather than trying to obtain some optimal and often unachievable technique exclusive to each specific problem. We have tried, by structuring our approach, starting with the output classification that the researchers are dealing with, to help the researcher into DSSs to get directly to the

type of tests they require rather than searching through many texts to try to achieve the same aim. In other words, enabling and saving time on performance evaluation and making it as simple as possible, whilst being appropriate and sufficient. We have started with the taxonomy of the types of output from NNs as a typical example of DSSs and how others have attempted to assess their performances, then led on to suggestions of how to deal with comparing and evaluating these. These suggestions use standard methods, but no one to our knowledge, has organised them in this overview manner to give very specific guidance as to how to go about testing the inherent performance of intelligent systems. These tests are what colleagues in artificial intelligence inform us they require to overcome their lack of statistical knowledge. Also, the statistics are organised in a readily accessible form for quick reference and made understandable to DSS researchers and developers. Since only these have thorough knowledge of their methods, the application of the above techniques requires careful consideration as to what is uniquely appropriate in terms of comparisons and tests for a particular set of data. Those who require more in-depth treatment of the subject matter will find this in more specialist literature in their field. Many texts have been written on individual parts or aspects of classification performance evaluation, but not an overall view.

Section 7 has suggested current methodologies for comparing the inherent performance of NN classifiers. This could be expanded to include other intelligent DSSs, as suggested by Hart and Wyatt [44]. It is hoped that we can contribute to the debate that will proceed towards realistic and achievable guidelines. It is anticipated that the current paper will help developers of intelligent DSSs to indicate that a scientific approach has been utilised in assessing the outputs. We have confined ourselves to empirical testing of the direct outputs from systems, not promoting the wider issues of evaluation. It is hoped that the guidelines for these can be formed by a committee of experts in the field, in the form of a white paper on this subject. We have aimed to only give a minimal sufficiency of inherent performance evaluation that is achievable by the DSS researchers and designers themselves, as part of this. The intention has been to give specific advice of a practicable, easy look-up nature to allow adequate performance evaluation, which may enable the plethora of systems that have been described in the literature to be turned into reality and applied in the medical domain, rather than just remaining on the shelf.

Acknowledgements

This research has been carried out as part of a Fellowship, funded by the Medical Research Council, London, UK.

Appendix A. The proforma used for Medline summaries

(a)	*Title*
(b)	*Date of publication*
(c)	*Intention*—applied or theoretical
(d)	*Gold standard*

Appendix A. (*Continued*)

(e)	*Degree of evaluation*
0	None
1	Train and test, cross-validation, etc.
1	Against gold standard, insufficient
1	Against invalid other methodology
2	Against gold standard, sufficient
3	Against another system, insufficient
4	Two systems, via gold standard, but insufficient, MAEs, etc.
5	Two systems, via gold standard, sufficient, sensitivity/specificity or gold standard not ideal
6	Full two systems, via gold standard, sufficient, full ROCs or other
(f)	*Classification type*
0	Clustering
1	Binary
2	Multiple
3	Continuous

References

[1] Aitkin M, Anderson DA, Francis B, Hinde HP. The fitting of exponential, Weibull and extreme value distributions to censored survival data using GLIM. Appl Stat 1980;29:156–63.

[2] Altman DG, Bland JM. The normal distribution. Br Med J 1995;310:298.

[3] Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991.

[4] Anand SS, Smith AE, Hamilton PW, Anand JS, Hughes JS, Bartel P. An evaluation of intelligent prognostic systems for colorectal cancer. Artif Intell Med 1999;15:105–19.

[5] Anderer P, Saletu B, Klöppel B, Semlitsch MV, Werner H. Discrimination between demented patients and normals based on topographic EEG slow-wave activity comparisons between *Z*-statistics, discriminant analysis and artificial neural network classifiers. Electroencephalogr Clin Neuropsychol 1996;91:108–17.

[6] Aprile D. Unionisation in a comparative neural network model: a trade union membership prediction in 12 states. Subst Use Misuse 1998;33(3):819–36.

[7] Arana E, Marti-Bonmati L, Oarades R, Bautista D. Focal calvarial bone lesions. Comparison of logistic regression and neural network models. Invest Radiol 1998;33:738–45.

[8] Astion ML, Wener MH, Thomas RG, Hunder GG, Bloch DA. Application of neural networks to the classification of giant cell arthritis. Arthritis Rheum 1994;37(5):760–7.

[9] Azuaje F, Dubitzky W, Lopes P, Black N, Adamson K, Wu X, et al. Predicting coronary disease risk based on short-term RR interval measurements: a neural network approach. Artif Intell Med 1999;15:275–97.

[10] Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991;115(11):843–8.

[11] Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in testing performance evaluation. Arch Pathol Lab Med 1986;35:13–20.

[12] Biagiotti R, Desii C, Vanzi E, Gacci G. Predicting ovarian malignancy: application of artificial neural networks to transvaginal and color Doppler flow US. Radiology 1999;210(2):399–403.

[13] Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. Stat Med 1998;17:1169–86.

[14] Boone JM, Seshagiri S, Steiner RM. Recognition of chest radiograph orientation for picture archiving and communications systems display using neural networks. J Digit Imaging 1992;5(3):190–3.
[15] Brouwer RK, MacAuley C. Classifying cervical cells using a recurrent neural network by building basins of attraction. Anal Quant Cytol Histol 1995;17(3):197–203.
[16] Buller D, Buller A, Innocent PR, Pawlak W. Determining and classifying the region of interest in ultrasonic images of the breast using neural networks. Artif Intell Med 1996;8(1):53–66.
[17] Bullinaria JA. Modeling reading, spelling, and past tense learning with artificial neural networks. Brain Lang 1997;59(2):236–66.
[18] Campbell JP, Maxey VA, Watson WA. The Hawthorne effect: implications for pre hospital research. Ann Emerg Med 1995;26:590–4.
[19] Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. Br Med J 2000;321:694–6, and http://www.mrc.ac.uk/complex_packages.html.
[20] Campbell MJ, Machin D. Medical statistics: a common sense approach. New York: Wiley; 1993.
[21] Chandra R, Optican LM. Detection, classification, and superposition resolution of action potentials in multiunit single-channel recordings by an on-line real-time neural network. IEEE Trans Biomed Eng 1997;44(5):403–12.
[22] Chiu C, Shanblatt MA. Human-like dynamic programming neural networks for dynamic time warping speech recognition. Int J Neural Syst 1995;6(1):79–89.
[23] Cohen IL, Sudhalter V, Landon-Jimenez D, Keogh M. A neural network approach to the classification of autism. J Autism Dev Disord 1993;23(3):443–66.
[24] Collett D. Modelling survival data in medical research. London: Chapman & Hall; 1997.
[25] Cox DR. Regression models and life tables. J R Stat Soc 1972;34(2):187–220.
[26] Daly F, Hand DJ, Jones MC, Lunn AD, NcConnway KJ. Elements of statistics. Avon: The Bath Press; 1995. p. 192–204.
[27] de Laurentis M, Ravdin PM. A technique for using neural network analysis to perform survival analysis of censored data. Cancer Lett 1996;77:127–38.
[28] Deligdisch L, Einstein AJ, Guera D, Gil J. Ovarian dysplasia in epithelial inclusion cysts. A morphometric approach using neural networks. Cancer 1995;76(6):1027–34.
[29] Delucci KL. The use and misuse of chi-square: Lewis and Burke re-visited. Psychol Bull 1983;94: 166–76.
[30] Derksen S, Keselman HJ. Backward forwards and stepwise automated selection algorithms. Frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 1992;45:265–8.
[31] Devoe LD. Computerized fetal heart rate analysis and neural networks in antepartum fetal surveillance. Curr Opin Obstet Gynecol 1996;8(2):119–22.
[32] Doig GS, Inman KJ, Sibbald WJ, Martin CM, Robertson JMcD. Modelling mortality in the intensive care unit. Comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. Proc Annu Symp Comput Appl Med Care 1993:361–5.
[33] Dowie J. The evaluation of decision aids: the role of the decision owner. Med Inform 1990;15:219–28.
[34] Dybowski R, Weller P, Chang R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 1996;347:1146–50.
[35] Ebell MH. Artificial neural networks for predicting failure to survive following in-hospital cardiopulmonary resuscitation. J Fam Pract 1993;36(3):297–303.
[36] Ely JW, Dawson JD, Mehr DR, Burns TL. Understanding logistic regression analysis through example. Fam Med 1996;28:134–40.
[37] Engelrecht R, Rector A, Moser W. Verification and validation. In: van Gennip EMSJ, Talmo JL, editors. Assessment and evaluation of information technologies. Amsterdam: IOS Press; 1995. p. 51–66.
[38] Ergemont-Petersen M, Talmon JL, Brender J, McNair P. On the quality of neural nets classifiers. Artif Intell Med 1994;6(5):359–81.
[39] Ferran EA, Ferrara P. Clustering proteins into families using artificial neural networks. Comput Appl Biosci 1992;8(1):39–44.
[40] Fischer H, Hennig J. Neural network-based analysis of MR time series. Magn Reson Med 1999;41(1):124–31.
[41] Friedman CP, Wyatt JC. Evaluation methods in medical informatics. New York: Springer-Verlag; 1997.

[42] Guiraud D. Application of an artificial neural network to the control of an active external orthosis of the lower limb. Med Biol Eng Comput 1994;32(6):610–4.
[43] Hanley JA, McNeil BJ. A method for comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
[44] Hart A, Wyatt J. Black boxes as medical decision aids: issues arising from a study of neural networks. Med Inform 1990;15(3):229–36.
[45] Heathfield HD, Pitty D, Hanka R. Evaluating information technology in healthcare: barriers and challenges. Br Med J 1998;316(7149):1959–61.
[46] Hilden J, Habbema JDF. Evaluation of clinical decision aids—more to think about. Med Inform 1990;15(3):275–84.
[47] Hornberger J, Goldstein MK. Clinical decision support systems: evaluating the evaluation. Med Decis Making 2000;20:130–1.
[48] Hunter A, Kennedy L, Henry J, Ferguson I. Application of neural networks and sensitivity analysis to improved prediction of trauma survival. Comput Methods Programs Biomed 2000;62:11–9.
[49] In Vitro Diagnostic Medical Devices Directive 98/79/EC. Off J Eur Communities 1998;L331:1.
[50] James CJ, Jones RD, Bones PJ, Carrol GJ. Detection of epileptiform discharges in the EEG by a hybrid system comprising mimetic, self-organized artificial neural network, and fuzzy logic stages. Clin Neurophysiol 1999;110(12):2049–63.
[51] Jefferson MF, Pendleton N, Lucas CP, Lucas SB, Horan MA. Evolution of artificial neural network architecture: prediction of depression after mania. Methods Inform Med 1998;37(3):220–5.
[52] Jefferson MF, Pendleton N, Lucas N, Horan MA. Neural networks (letter). Lancet 1995;346:1712.
[53] Kaplan B. Addressing organisational issues into the evaluation of medical systems. J Am Med Inform Assoc 1997;4:94–110.
[54] Kaplan EI, Mieir P. Non-parametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.
[55] Karakitsos P, Stergiou EB, Pouliakis A, Tzivras M, Archimandritis A, Liossi AI, et al. Potential of the back propagation neural network in the discrimination of benign from malignant gastric cells. Anal Quant Cytol Histol 1996;18(3):245–50.
[56] Katz RT, Campagnolo DI, Goldberg G, Parker JC, Pine ZM, Whyte J. Critical evaluation of clinical research. Arch Phys Med Rehabil 1995;76:82–93.
[57] Koss LG, Sherman ME, Cohen MB, Anes AR, Darragh TM, Lemos LB, et al. Significant reduction in the rate of false-negative cervical smears with neural network-based technology (PAPNET testing system). Hum Pathol 1997;28(10):1196–203.
[58] Kosugi Y, Sase M, Suganami Y, Uemoto N, Momose T, Nishikawa J. Neural network-based PET image reconstruction. Methods Inform Med 1997;36(4–5):329–31.
[59] Kumar DK, Pah ND. Neural networks and wavelet decomposition for classification of surface electromyography. Electromyogr Clin Neurophysiol 2000;40(7):411–21.
[60] Lapuerta P, Azen SP, LaBree L. Use of neural networks in predicting the risk of coronary artery disease. Comput Biomed Res 1995;28:38–52.
[61] Lemeshow S, Teres D, Klar J. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. J Am Med Assoc 1993;270:2478–86.
[62] Leong PH, Jabri MA. Kakadu—a low power analogue neural network classifier. Int J Neural Syst 1993;4(4):381–94.
[63] Lette J, Colletti BW, Cerino M, McNamara D, Eybalin M-C, Levasseur A, et al. Artificial intelligence versus logistic regression statistical modelling to predict cardiac complications after non-cardiac surgery. Clin Cardiol 1994;17:609–14.
[64] Lin JS, Cheng KS, Mao CW. Multispectral magnetic resonance images segmentation using fuzzy Hopfield neural network. Int J Biomed Comput 1996;42(3):205–14.
[65] Lisboa PJG, Ifeachor EC, Szczzepaniak PS, editors. Artificial neural networks in biomedicine. Heidelberg: Springer-Verlag; 2000.
[66] Lloyd-Williams M, Williams TS. A neural network approach to analyzing health care information. Top Health Inform Manage 1996;17(2):26–33.
[67] Lohmann R, Schneider G, Behrens D, Wrede P. A neural network model for the prediction of membrane-spanning amino acid sequences. Protein Sci 1994;3(9):1597–601.

[68] Marble RP, Healy JC. A neural network approach to the diagnosis of morbidity outcomes in trauma care. Artif Intell Med 1999;15:299–307.
[69] Mariak Z, Swiercz M, Krejza J, Lewko J, Lyson T. Intracranial pressure processing with artificial neural networks: classification of signal properties. Acta Neurochir (Wien) 2000;142(4):407–12.
[70] Mariani L, Coradini D, Bugnazoli E, Boracchi P, Marubini E, Pilotti S, et al. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox's regression model and its artificial neural network extension. Breast Cancer Res Treat 1997;44:167–78.
[71] Mazzone P, Fortuna L, Arena P, Pisani R. Multi-layer neural network analysis of cerebrospinal fluid pressure patterns in idiopathic normal-pressure hydrocephalus. Technol Health Care 1996;4(4):393–401.
[72] Medical Devices Directive 93/42/EC. Off J Eur Communities 1993;L139:1.
[73] Meting HJ, Coenegracht MJ. Neural networks in high-performance chromatography optimization: response surface modelling. J Chromatogr 1996;728:47–53.
[74] Michie D, Spiegelhalter DJ, Taylor CC, editors. Machine learning, neural and statistical classification. 2001. Statlog website http://www.ncc.up.pt/liacc/ml/statlog/.
[75] Miller PL, Sittig DF. The evaluation of clinical decision support systems: what is necessary versus what is interesting. Med Inform 1990;15(3):185–90.
[76] Mobley BA, Schechter E, Moore WE, McKee PA, Eichner JE. Predictions of coronary artery stenosis by artificial neural network. Artif Intell Med 2000;18(3):187–203.
[77] Nohr C. The evaluation of expert diagnostic systems. How to assess outcomes and quality parameters. Artif Intell Med 1994;6:123–35.
[78] Nugent CD, Lopez J, Smith AE, Black ND. Prediction models in the design of neural network-based ECG classifiers: a neural network and genetic programming approach. BMC Med Inform Decis Making 2002;2:1.
[79] Nugent CD, Webb JAC, Black ND, Wright GTH, McIntyre M. An intelligent framework for the classification of the 12 lead ECG. Artif Intell Med 1999;16:3–23.
[80] O'Moore R, Englebrecht R. The evaluation of medical decision support and expert systems: reflections on the literature. In: Lecture notes in medical informatics. New York: Springer-Verlag; 1991. p. 263–73.
[81] Ohno-Machado L. A comparison of Cox's proportional hazards and artificial neural network models for medical prognosis. Comput Biol Med 1997;27:55–65.
[82] Ortiz J, Ghefter CG, Silva CE, Sabbatini RM. One-year mortality prognosis in heart failure: a neural network approach based on echocardiographic data. J Am Coll Cardiol 1995;26(7):1586–93.
[83] Park HA, Lee EO, Song MS. Development of a nursing diagnosis system using a back-propagation neural network model: an application for stomach cancer patients. Medinfo 1995;8(2):1399–403.
[84] Patel MM, Rayburn DB, Browning JA, Kline JA. Neural network analysis of the volumetric capnogram to detect pulmonary embolism. Chest 1999;116(5):1325–32.
[85] Perneger TV. What's wrong with Bonferroni adjustments. Br Med J 1998;316:1236–8.
[86] Reggia J. Evaluation of Medical Expert Systems. In: Proceedings of the Symposium on Computer-Assisted Medicine and Decision-Making, A Case Study in Performance Assessment. 1985. p. 287–329.
[87] Rehman HU, Linkens DA, Asbury AJ. Neural networks and non-linear regression modelling and control of depth of anaesthesia for spontaneously breathing and ventilated patients. Comput Methods Programs Biomed 1993;40(4):227–47.
[88] Reid JC, Nair SS, Kashani JH, Rao VG. Detecting dysfunctional behavior in adolescents: the examination of relationships using neural networks. Proc Annu Symp Comput Appl Med Care 1994;743–6.
[89] Ronco AL. Use of artificial neural networks in modelling associations of discriminant factors: towards and intelligent selective breast cancer screening. Artif Intell Med 1999;16:299–309.
[90] Rossi-Mori A, Pisanelli DM, Ricci F. Evaluation stages and design steps for knowledge-based systems in medicine. Med Inform 1990;15(3):191–204.
[91] Rubegni P, Cevenini G, Flori ML, Barbini P, Andreassi L. Relationship between minimal phototoxic dose and skin colour plus sun exposure history: a neural network approach. Photodermatol Photoimmunol Photomed 1998;14(1):26–30.
[92] Scott JA. Using artificial neural network analysis of global ventilation-perfusion scan morphometry as a diagnostic tool. AJR Am J Roentgenol 1999;173(4):943–8.
[93] Sepulveda F, Wells DM, Vaughan CL. A neural network representation of electromyography and joint dynamics in human gait. J Biomech 1993;26(2):101–9.

[94] Shortliffe EH, Davis R. Some considerations for the implementation of knowledge-based expert systems. SIGART Newslett 1975;35:9–12.
[95] Silipo R, Gori M, Taddei A, Varanini M, Marchesi C. Classification of arrhythmic events in ambulatory electrocardiogram, using artificial neural networks. Comput Biomed Res 1995;28(4):305–18.
[96] Smith AE, Anand SS. Patient survival estimation with multiple variables: adaptation of Cox's regression to give an individual's point prediction. In: Proceedings of the IDAMAP. Berlin, 2000. p. 51–4.
[97] Smith AE, McClean SI, Nugent CD. Towards sufficiency of performance evaluation for intelligent systems in medicine. In: Proceedings of the IFMBE. vol. 2. Pula, Croatia, 2001. p. 1102–4.
[98] SPSS regression models. V10.0. Chicago: SPSS Inc.; 1999.
[99] Sveinsson JR, Benediktsson JA, Stefansson SB, Davidsson K. Parallel principal component neural networks for classification of event-related potential waveforms. Med Eng Phys 1997;19(1):15–20.
[100] Swiercz M, Mariak Z, Lewko J, Chojnacki K, Kozlowski A, Piekarski P. Neural network technique for detecting emergency states in neurosurgical patients. Med Biol Eng Comput 1998;36(6):717–22.
[101] Tafeit E, Moller R, Sudi K, Reibnegger G. Artificial neural networks compared to factor analysis for low-dimensional classification of high-dimensional body fat topography data of healthy and diabetic subjects. Comput Biomed Res 2000;33(5):365–74.
[102] Tian J, Juhola M, Gronfors T. Related articles latency estimation of auditory brainstem response by neural networks. Artif Intell Med 1997;10(2):115–28.
[103] Tourassi GD, Floyd CE, Coleman RE. Acute pulmonary embolism: cost-effectiveness analysis of the effect of artificial neural networks on patient care. Radiology 1998;206(1):81–8.
[104] Tsujii O, Freedman MT, Mun SK. Automated segmentation of anatomic regions in chest radiographs using an adaptive-sized hybrid neural network. Med Phys 1998;25(6):998–1007.
[105] Vach W, RoBner R, Schumacher M. Neural networks and logistic regression. Part II. Comput Stat Data Anal 1996;21:683–701.
[106] van Bemmel JH, Musen MA. Handbook of medical informatics. Website v3.3. 2001. http://www.mieur.nl/mihandbook/r_3_3/handbook/home.htm.
[107] Van Hoey G, De Clercq J, Vanrumste B, Van De Walle R, Lemahieu I, D'Have M, et al. EEG dipole source localization using artificial neural networks. Phys Med Biol 2000;45(4):997–1011.
[108] Viktor HL, Cloete I, Beyers N. Extraction of rules for tuberculosis diagnosis using an artificial neural network. Methods Inform Med 1997;36(2):160–2.
[109] Wang L, Ross J. Variable threshold as a model for selective attention, (de)sensitization, and anesthesia in associative neural networks. Biol Cybern 1991;64(3):231–41.
[110] Wright IA, Gough NA. Artificial neural network analysis of common femoral artery Doppler shift signals: classification of proximal disease. Ultrasound Med Biol 1999;25(5):735–43.
[111] Wyatt JC. Evaluation of clinical information systems. In: van Bemmel JH, Musen MA, editors. Handbook of medical informatics. Heidelberg: Springer-Verlag; 1997. p. 463–9.
[112] Zaharia CN, Cristea A. A micropopulational modelling of a viral epidemic by using a special neural network. Stud Health Technol Inform 1999;68:682–5.
[113] Zweig MH, Campbell G. Receiver-operating characteristics (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39(4):561–77.