Artificial neural networks are able to recognize gastro-oesophageal reflux disease patients solely on the basis of clinical data

Fabio Pace[a], Massimo Buscema[b], Patrizia Dominici[c], Marco Intraligi[b], Fabio Baldi[d], Renzo Cestari[e], Sandro Passaretti[f], Gabriele Bianchi Porro[a] and Enzo Grossi[c]

Background **Artificial neural networks (ANN) are modelling mechanisms that are highly flexible and adaptive to solve the non-linearity inherent in the relationship between symptoms and underlying pathology.**

Objectives **To assess the efficacy of ANN in achieving a diagnosis of gastro-oesophageal reflux disease (GORD) using oesophagoscopy or pH-metry as a diagnostic gold standard and discriminant analysis as a statistical comparator technique in a group of patients with typical GORD symptoms and with or without GORD objective findings (e.g. a positive oesophagoscopy or a pathological oesophageal pH-metry).**

Methods **The sample of 159 cases (88 men, 71 women) presenting with typical symptoms of GORD, were subdivided on the basis of endoscopy and pH-metry results into two groups: GORD patients with or without oesophagitis, group 1 ($N=103$), and pH and endoscopy-negative patients in whom both examinations were negative, group 2 ($N=56$). A total of 101 different independent variables were collected: demographic information, medical history, generic health state and lifestyle, intensity and frequency of typical and atypical symptoms based on the Italian version of the Gastroesophageal Reflux Questionnaire (Mayo Clinic). The diagnosis was used as a dependent variable. Different ANN models were assessed.**

Results **Specific evolutionary algorithms selected 45 independent variables, concerning clinical and demographic features, as predictors of the diagnosis. The highest predictive performance was achieved by a 'back propagation' ANN, which was consistently 100% accurate in identifying the correct diagnosis compared with 78% obtained by traditional discriminant analysis.**

Conclusion **On the basis of this preliminary work, the use of ANN seems to be a promising approach for predicting diagnosis without the need for invasive diagnostic methods in patients suffering from GORD symptoms. *Eur J Gastroenterol Hepatol* 17:605–610 © 2005 Lippincott Williams & Wilkins.**

European Journal of Gastroenterology & Hepatology 2005, **17**:605–610

Keywords: artificial neural networks, gastro-oesophageal reflux disease, gastro-oesophageal reflux disease symptoms, non-gastro-oesophageal reflux disease, pH-metry

[a]Department of Gastroenterology, L. Sacco H., Milan, Italy, [b]Semeion Research Center of Sciences of Communication, Rome, Italy, [c]Medical Department, Bracco Imaging SpA, Milan, Italy, [d]Department of Gastroenterology, S. Orsola Malpighi H., Bologna, Italy, [e]Department of Gastroenterology, University of Brescia, Brescia, Italy and [f]Department of Gastroenterology, S. Raffaele H., Milan, Italy.

Correspondence to Fabio Pace, MD, Department of Gastroenterology, L. Sacco University Hospital, Via G.B. Grassi 74, 20157 Milan, Italy.
Tel: +39 02 39042486 385; fax: +39 02 39042232;
e-mail: cn.fapac@tin.it

Introduction

Gastro-oesophageal reflux disease (GORD) has recently been defined as the presence of oesophageal mucosal interruptions or by the occurrence of reflux-induced symptoms severe enough to impair quality of life significantly [1]. The importance of symptom evaluation and assessment is already evident from the definition in the management of the disease. Despite the fact that symptom analysis is considered to be important to determine whether reflux-induced symptoms are sufficiently severe to justify the diagnosis of reflux disease, there is no reliable method to use this kind of information as a diagnostic tool, as a result of the non-linearity of the relationship between the quality and intensity of symptoms and the diagnostic target.

Artificial neural networks (ANN) represent a novel algorithmic approach for the solution of non-linear problems that are too complex for conventional statistic analysis [2–6].

The aim of this study was to ascertain whether ANN can be used in a population with symptoms suggestive of GORD to discriminate between patients with or without oesophagitis, but with a pathological oesophageal pH condition, and individuals with neither oesophagitis nor pathological reflux, on the basis of symptoms analysis solely.

Materials and methods

As summarized in Table 1, the study considered 159 subjects (88 men, 71 women) referred to four

0954-691X © 2005 Lippincott Williams & Wilkins

Table 1 Population in study

Subjects, *N*	159
Male (%)	88 (55·3)
Female (%)	71 (44·7)
Mayo Clinic questionnaire administered (%)	159 (100)
Oesophagoscopy (%)	159 (100)
With oesophagitis (%)	62 (39)
Without oesophagitis (%)	97 (61)
pH-Metry 24 h (%)	97 (61)
Pathological gastro-oesophageal reflux (%)	41 (25)
Normal gastro-oesophageal reflux (%)	56 (36)

university gastroenterology centres (Bologna, Brescia, Milan San Raffaele and Milan L. Sacco) presenting with symptoms suggestive of GORD undergoing an upper gastrointestinal endoscopy; all subjects with a negative endoscopy subsequently underwent 24-h oesophageal pH monitoring.

Local Ethics Committees approved the study and all patients were asked to give their written informed consent.

All patients were assessed by means of the Gastro-Esophageal Reflux Questionnaire proposed by the Mayo Clinic [7], which has been adapted to be used in the Italian language. Linguistic validation of the Italian version was obtained through a formal translation and back-translation process. The Gastro-Esophageal Reflux Questionnaire is a 96-item questionnaire, which explores not only the present symptoms of the patient, but also demographic, social and other clinical characteristics, based on questions relating to 'typical' and 'atypical' GORD symptoms.

On the basis of the two instrumental tests, the group was divided as follows: 103 GORD patients (62 with oesophagitis grades 1–4 according to Savary and 41 without oesophagitis but with a pathological oesophageal pH monitoring, i.e. with a duration of oesophageal pH $< 4{\cdot}0$ greater than 5% of the monitoring period), group 1 and 56 pH and endoscopy-negative subjects (no oesophagitis, no abnormal reflux), group 2.

Artificial neural network structure and architecture

ANN models were constructed by the use of non-commercial programs developed by the Semeion Research Center [8].

Four different ANN architectures with two different learning rules were assessed in this experiment: feed-forward back propagation [8], sine net learning law [9], auto-recurrent [10], and cluster-recurrent [11].

All ANN had the following structure: the input vector had a number of nodes equal to the number of independent variables; the output vector had two nodes corresponding to the two different diagnoses (group 1 versus group 2); one layer of hidden units.

Cross-validation research protocol

The validation protocol consisted of the following steps: (i) Subdividing the database in a random way into two subsamples: the first named the training set and the second called the testing set; (ii) Choosing a fixed ANN (or organism) that is trained on the training set. In this phase the ANN learns to associate the input variables with those indicated as targets; (iii) At the end of the training phase, the weight matrix produced by the ANN is saved and 'frozen' together with all of the other parameters used for the training; (iv) The testing set, which has not yet been used, is shown to the ANN, which can make its own evaluation, based on the previously carried out training. This operation takes place for each input vector and every result (output vector) is not communicated to the ANN; (v) The ANN is in this way evaluated only with reference to the generalization ability that it has acquired during the training phase; (vi) A new ANN is constructed with identical architecture to the previous one and the procedure is repeated from point (i).

In the first phase, n copies of samples, with which $n \times 2$ elaborations were carried out, were normally generated in a random way from the global database. The elaborations were created by inverting the pairs of samples, i.e. by using each sample once for the training and once for the testing. This procedure allows a 'non-polarized evaluator' to be obtained. This phase of the protocol is illustrated in the left diagram of Figure 1.

Database optimization for neural networks

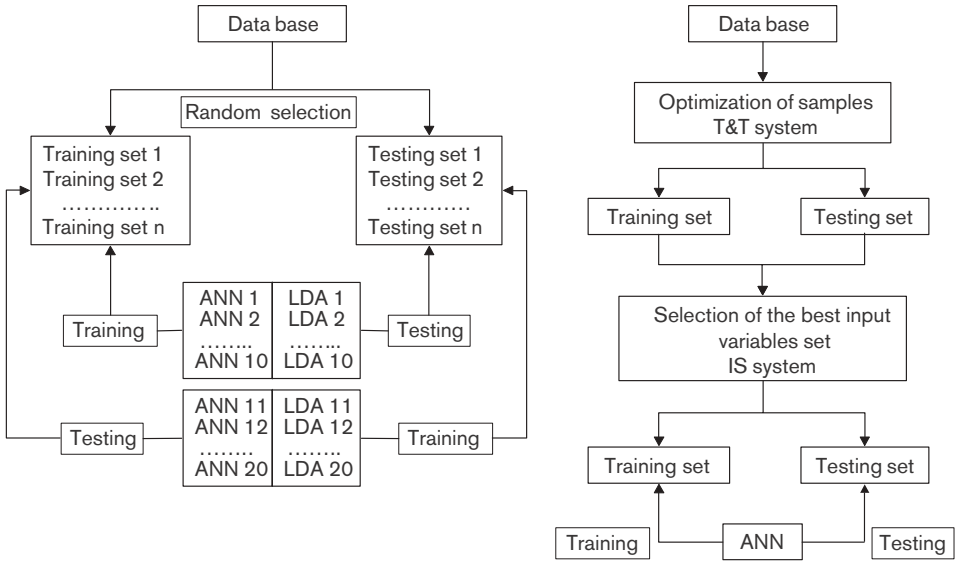
In order to optimize the training and testing (T&T) procedure and to reduce the number of the input variables, selecting the most relevant ones for the target prediction irrespective of the value of their linear correlation with the dependent variable, the T&T model and the input selection (IS) system, both originally developed by Semeion [11], were used respectively.

The T&T system is a population of ANN finalized from the training of an optimum model of distribution of the complete dataset into a T&T subset (Fig. 1).

T&T is based on an evolutionary algorithm developed by the Semeion Research Center, the genetic dopping algorithm (GenD), which has demonstrated greater efficiency with respect to the limits of the classic genetic algorithms in resolving complex optimization problems.

The optimization algorithm distributes the original sample in two or more subsamples to obtain the maximum performance possible on behalf of an ANN

Fig. 1



Cross-validation research protocol and database optimization for neural networks.
First, the dataset is randomly subdivided into two equal parts (left part of the figure). One part is used as a training set, in which the net learns knowing the dependent variable, and the other part is used as a testing set, in which the net is asked to classify each record without knowing the dependent variable, then vice versa. This cross-validation procedure is repeated 10 times on different random subsamples. Therefore, on the whole, both for neural networks and for discriminant analysis, 20 cross-validation procedures are performed. In the second part of the method (right part of the figure) the so-called training and testing (T&T) system optimizes the two subsamples and the input variables are selected by the input selection (IS) system, then networks are used (see text for futher details).
ANN, Artificial neural networks; LDA, linear discriminant analysis.

that is trained on the first sample and validated on the second. The score reached by each ANN represents its fitness, and its probability of evolution. It is possible to limit eventual optimistic polarizations in the evaluation of the performances, to invert the two samples, and to consider the average between the two estimates obtained as the fitness of the algorithm and an estimate of the quality of the model.

The limit of the optimistic estimate of the model's performances is fundamentally caused by the presence of errors in the database, or the presence of areas that are not sufficiently represented in the original sample.

The substantial difference between this approach and a traditional approach, which is based on the random subdivision of the original sample, is the possibility of taking advantage of all the information present in the dataset.

The IS system is a selection mechanism of the input variables of a fixed dataset, based on the evolutionary algorithm GenD. As data are collected to build the dataset, the relationships between the variables detected and the function of the process under examination are not known, therefore a subset of variables on which to build the model is selected. Data collection is performed by trying to include all variables that can have a connection with the event under investigation. As a result, often a series of variables that do not contain any information regarding the process being examined is present. When inserted into the model, these variables cause an increase of noise and greater difficulty for the ANN to learn data correctly. Consequently, a heuristic approach, which neglects other types of values and concentrates on a set of input data that provide the best performance in an ANN training model, was chosen.

Given their nature, the T&T and IS systems can be viewed as data preprocessing organisms with a high capacity to single out the information useful to optimize the relationship between the input and output variables during the calculation process. To obtain this result, ANN populations that evolve, according to the criteria of the GenD evolutionary algorithm, are used to reach the best representation of variables and of the sample size of the problem to be resolved.

Data analysis

Two different experiments were planned following an identical research protocol. The first included all 101 independent variables extracted from the questionnaire,

Table 2 Best performances of the various experiments with artificial neural networks

	GORD		Normal		Global testing		
	Errors	% Accuracy	Errors	% Accuracy	Errors	AA	PA
Database 101 variables	3	85.00	9	67.86	12	73.43	75.00
Database 45 variables	0	100.00	0	100.00	0	100.00	100.00
LDA	12	78.05	9	78.57	21	78.31	78.35

GORD, Gastro-oesophageal reflux disease. Errors: bad classifications – accuracy: percentage of correct classifications – arithmetic average (AA) – ponderate average (PA) versus linear discriminant analysis (LDA).

Table 3 Forty-five variables selected by the input selection system and corresponding input relevance for artificial neural networks

Heartburn	4306	Persistent gastric/intestinal pain	2084
Hiatus hernia	3430	Cough intensity	2069
Prescribed examinations for GORD	3369	Frequency of chest pain	2029
Periodic frequency of swallowing problems	3263	Waking at night–retrosternal pain	2010
Aspirin, frequency of use	3219	Antirheumatic drugs, frequency of use	2009
Intensity of swallowing problems	3204	Pneumonia	1959
Frequency of medical check-ups	3091	Cough	1831
Variation in weight in past year	2783	Oesophageal dilation	1830
Coffee drinker	2735	Heart disorders	1825
Retrosternal pain	2732	Regular smoker	1781
Vomiting (frequency)	2669	Acid reflux in mouth	1766
Lump in throat	2605	Interference with daily activities	1630
Relative with gastroduodenal disorder	2558	Marital status	1603
Cough frequency/year	2501	Slow walk–chest pain	1544
Heart therapy	2488	Intensity of chest pain	1455
Chest pain in past year	2445	Alcohol units/week in past year	1437
Medical visits for GORD	2383	Episodes of breathlessness	1433
Belching	2359	Asthma	1238
Sibilant rhonchi or wheezing	2270	Cough at night	1198
Ingestion of beverages–chest pain	2235	Total acid reflux	1185
Sleeping semisupine	2211	Oesophageal surgery	1148
Health state during past year	2168	Hiccups	839
Intensity of acid reflux	2136		

GORD, Gastro-oesophageal reflux disease.

including the frequency and intensity of typical and atypical GORD symptoms, plus numerous other social and demographic characteristics, clinical features and history. In the second experiment, the IS system coupled with the T&T system automatically selected the most relevant variables, and therefore 45 variables were included in the model.

Discriminant analysis

Discriminant analysis was also performed on the same datasets to evaluate the predictive performance of this advanced statistical method by a statistician blinded to the ANN results. Different models were assessed to optimize the predictive ability. In each experiment a different number of variables was included and the sample was randomly divided into two subsamples, one for the training phase and the other for the testing phase, with the same record distributions used for ANN validation.

Results

The conventional, even though advanced, statistical analysis, which is based on the assumption of linear correlations, revealed a limited capacity for discriminating between patients of group 1 and group 2, i.e. patients with or without pathological endoscopy or pH monitoring, respectively. Discriminant analysis achieved 78% accuracy (Table 2). Table 2 also summarizes the results obtained using the neural networks: using all 101 independent variables the predictive accuracy was 85% in selecting GORD patients, but only 67.9% in selecting pH and endoscopy-negative subjects, which was by no means satisfactory.

In a subsequent step, thanks to the IS system coupled with T&T system (see right part of Fig. 1), which selected 45 variables, the best net reached a predictive accuracy of 100%, i.e. it was able to predict the diagnosis of objectively confirmed GORD without error (see Table 2). Table 3 lists the 45 variables selected by the IS system. The predictive accuracy values become stable thereafter, and did not vary when neural networks of the same type were used consecutively to repeat the analysis.

Discussion

Our study is the first, to our knowledge, to aim at achieving a diagnosis of GORD by means of the application of ANN. The use of ANN has recently been proposed in many different fields, from cardiology [12] to neurology [13], dermatology [14], pneumology [15] and many others. Interest is also growing in the gastroenterology field, as witnessed by recent editorials [16] and

articles [17,18]. In our study, aimed at assessing the presence of GORD, ANN were more efficient than conventional statistical analysis, e.g. discriminant analysis, in achieving a correct performance (diagnosis) in an extremely high percentage of subjects on average, up to a predictive accuracy of 100% when the best net was used. Diagnosing GORD by conventional invasive methods such as upper gastrointestinal endoscopy or oesophageal pH-metry has many limitations. These tests are cumbersome, costly, and moreover display limited sensitivity and specificity, as they are able to detect only a limited part of the GORD patient spectrum, i.e. the presence of mucosal damage or the presence of an excessive oesophageal acid exposure caused by GORD, but are insensitive in those patients with symptoms but without the above GORD complications, also referred to as non-GORD patients and functional heartburn patients [19]. A diagnosis based on symptoms is thus very appealing, because symptoms are the main reason that causes patients with GORD to seek medical attention. It is probably more important to address treatment to these patients than to those with the presence of some damage or excessive acid exposure in the distal oesophageal mucosa. Symptoms are, however, an elusive entity; they are frequently found in the general population, in up to 45% of Italians [20], i.e. more than three times the expected GORD prevalence. They can be divided into 'typical' or 'atypical', and astonishingly the latter can be found even more frequently than the former in the general population [21]. They can be volunteered by the patient or only elicited on questioning by a physician and the proportion of the two may vary notably [22]. Finally, they can overlap, at least partly, with symptoms of functional dyspepsia [23]. For this reason, it has been hypothesized that 'the use of short, self-administered questionnaires in routine clinical care will improve the reliability of the separation of reflux-induced symptoms from true dyspepsia' [24], and more generally, of GORD sufferers from normal individuals.

However, the state of the art of such questionnaires is still unsatisfactory, for many reasons; the principle one being the fact that none of those proposed so far has developed beyond the phase of initial validation or has been tested against the diagnosis of GORD achieved with 'robust', i.e. instrumental, tests such as endoscopy or pH-metry.

We have decided to build up a questionnaire starting by identifying, in our sample of GORD patients, a number of presenting clinical features, by means of a very sophisticated mathematical tool, e.g. the ANN. This approach has two advantages: first, it does not require an 'a priori' theorization of the complex function underlying the relationship between independent and dependent variables. Second, it overcomes the possibility of some features (independent variables) being correlated with each other or with the diagnosis in an unpredictable way. In other words, no a priori model is requested when ANN are applied, but rather the model is able to find *a posteriori* which ones are best correlated with the outcome. Furthermore, the possible reciprocal relationships do not negatively influence the performance of the neural networks.

By serial applications of the various ANN families we used, we were able to reduce the number of independent variables from the initial number of 101 (the number of items contained in the Mayo Clinic questionnaire [7] plus some basic demographic characteristics) to 45. Even if it seems that this number is too large, it must be remembered that the original questionnaire from the Mayo Clinic contained double the number of items. Second, contrary to normal statistical methods, such as logistic regression or discriminant analysis, which need a limited number of independent variables to be selected on the basis of the linear correlation with the dependent variable to predict outcome (diagnosis or prognosis), ANN are able to use all the information contained in the sample of variables, and are by no means affected by the non-linear correlation among independent variables. In other words, whereas the complex and often non-linear relationship among variables is a problem for traditional statistical methods and requires keeping to a minimum the number of independent variables, it is an advantage when ANN are used, and no reduction of the number of variables is needed.

The list of the 45 variables is displayed in Table 3; this list will provide the basis (process of item generation) for the new GORD questionnaire, to be subsequently evaluated in a prospective way in a clinical study, which is at present at an advanced stage of preparation.

Conflict of interest
None declared.

Authors' contributions
P.D. and E.G. proposed the study and developed the protocol. F.B., R.C., S.P. and F.P. recruited patients and performed the examinations. M.B. and M.I. performed the ANN analysis. F.P., E.G. and M.G. wrote the manuscript and G.B.P. was responsible for extensive revisions.

References

1 Dent J, Brun J, Fendrick AM, Fennerty MB, Janssens J, Kahrilas PJ, *et al.* An evidence-based appraisal of reflux disease management–the Genval Workshop Report. *Gut* 1999; **44 (Suppl. 2)**:S1–S16.
2 Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet* 1995; **346**:1075–1079.
3 Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet* 1995; **346**:1135–1138.
4 Buscema M. Self recurrent neural networks. *Substance Use Misuse* 1998; **33**:495–501.
5 Dayhoff JE, DeLeo JM. Artificial neural networks: opening the black box. *Cancer* 2001; **91**:1615–1635.
6 Hollander D. Is artificial intelligence an intelligence choice for gastroenterologist? *Dig Liv Dis* 2003; **35**:212–214.
7 Locke GR, Talley NJ, Weaver AL, Zinsmeister AR, *et al.* A new questionnaire for gastro-oesophageal reflux disease. *Mayo Clin Proc* 1994; **69**:539–547.

8 Buscema M, Sacco P. *Feed forward networks in financial prediction: the future that modifies the present*. Rome: Expert Systems; 2000.
9 Buscema M. *Sine net: a new learning rule for adaptive systems. Semeion technical paper 21*. Rome: Semeion Research Centre; 2000.
10 Buscema M. Artificial neural networks and complex social systems. *Substance Use Misuse* 1998; **33**:495–501.
11 Buscema M, and the Semeion Group. *Artificial neural networks and complex social systems*. Milan: Franco Angeli Publisher; 1999. pp. 452–457.
12 Olsson S, Ohlsson M, Ohlin H, Edenbrandt L. Neural networks – a diagnostic tool in acute myocardial infarction with concomitant left branch block. *Clin Physiol Funct Imag* 2002; **22**:295–299.
13 Castellaro C, Favaro G, Castellaro A, Casagrande A, Castellaro S, Puthenparampil DV, Salimbeni CF, *et al.* An artificial intelligence approach to classify and analyse EEG traces. *Neurophysiol Clin* 2002; **32**:193–214.
14 Piccolo D, Ferrari A, Peris K, Daidone R, Ruggeri B, Chimenti S. Dermoscopic diagnosis by a trained clinician vs a clinician with minimal dermoscopy trained vs computer-aided diagnosis of 341 pigmented skin lesion: a comparative study. *Br J Dermatol* 2002; **147**:481–486.
15 Bibi H, Nutman A, Shoseyov D, Shalon M, Peled R, Kivity S, Nutman J, *et al.* Prediction of emergency department visits for respiratory symptoms using an artificial neural network. *Chest* 2002; **122**:1627–1632.
16 Hollander D. Is artificial intelligence an intelligent choice for gastroenterologists? *Dig Liv Dis* 2003; **35**:212–214.
17 Andriulli A, Grossi E, Buscema M, Festa V, Intraligi NM, Dominici P, Cerutti P, Perri F, NUD LOOK Study Group *et al.* Contribution of artificial neural networks to the classification and treatment of patients with uninvestigated dyspepsia. *Dig Liv Dis* 2003; **35**:222–231.
18 Selaru FM, Xu Y, Yin J, Zou T, Liu TC, Mori Y, Abraham JM, Sato F, Wang S, Twigg C, Olaru A, Shustova V, Leytin A, Hytiroglou P, Shibata D, Harpaz N, Meltzer SJ, *et al.* Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. *Gastroenterology* 2002; **122**:606–613.
19 The Rome II International Working Teams. Functional heartburn. In: Drossman DA, Corazziari E, Talley NJ, *et al.* (editors): *Rome II: the functional gastrointestinal disorders*. 2nd ed. Lawrence: Allen Press Inc.; 2000, pp. 275–278.
20 Valle C, Broglia F, Pistorio A, Tinelli C, Perego M. Prevalence and impact of symptoms suggestive of gastroesophageal reflux disease. *Dig Dis Sci* 1999; **44**:1848–1852.
21 Ruth M, Mansson I, Sandberg N. The prevalence of symptoms suggestive of esopahgeal disorders. *Scand J Gastroenterol* 1991; **26**:73–81.
22 Jones RH, Pali A, Hungin S, Phillips J, Mills JG. Gastro-oesophageal reflux disease in primary care in Europe: clinical presentation and endoscopic findings. *Eur J Gen Pract* 1995; **1**:149–154.
23 Small PK, Loudon MA, Waldron B. Importance of reflux symptoms in functional dyspepsia. *Gut* 1995; **36**:189–192.
24 Dent J. Definition of reflux disease and its separation from dyspepsia. *Gut* 2002; **50 (Suppl IV)**:17–20.