Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 19 (2006) 408–415
Neural Networks
www.elsevier.com/locate/neunet

The use of artificial neural networks in decision support in cancer: A systematic review

Paulo J. Lisboa [a], Azzam F.G. Taktak [b,*]

[a] School of Computing and Mathematical Science, Liverpool John Moores University, Liverpool, UK
[b] Department of Clinical Engineering, Royal Liverpool University Hospital, 1st Floor, Duncan Building, Daulby Street, Liverpool L7 8XP, UK

Received 10 January 2005; accepted 31 October 2005

Abstract

Artificial neural networks have featured in a wide range of medical journals, often with promising results. This paper reports on a systematic review that was conducted to assess the benefit of artificial neural networks (ANNs) as decision making tools in the field of cancer. The number of clinical trials (CTs) and randomised controlled trials (RCTs) involving the use of ANNs in diagnosis and prognosis increased from 1 to 38 in the last decade. However, out of 396 studies involving the use of ANNs in cancer, only 27 were either CTs or RCTs. Out of these trials, 21 showed an increase in benefit to healthcare provision and 6 did not. None of these studies however showed a decrease in benefit. This paper reviews the clinical fields where neural network methods figure most prominently, the main algorithms featured, methodologies for model selection and the need for rigorous evaluation of results.

Keywords: Artificial neural networks; Decision support; Cancer; Clinical trials; Randomised-controlled trials

1. Introduction

In the last decade, the use of artificial intelligence (AI) has become widely accepted in medical applications. This is manifested by an increasing number of medical devices currently available on the market with embedded AI algorithms, together with an accelerating pace of publication in medical journals, with over 500 academic publications each year featuring Artificial Neural Networks (ANNs) (Gant, Rodway, & Wyatt 2001). Claimed advantages of neural network methods include:

- Ease of optimisation, resulting in cost-effective and flexible non-linear modelling of large data sets.
- Accuracy for predictive inference, with potential to support clinical decision making.
- These models can make knowledge dissemination easier by providing explanation, for instance, using rule extraction or sensitivity analysis (Lisboa, 2002).

The published literature suggests that ANN models have been shown to be valuable tools in reducing the workload on the clinicians by detecting artefact and providing decision support, potentially with the ability to automatically re-estimate the model on-line. However, there are relatively few published clinical trials, and even fewer testing the clinical value of ANNs against established linear-in-the-parameters statistical methods (Lisboa, 2002).

There are two recurring concerns on ANNs. The first is the use of first principle statistical methods to control model complexity, which has been addressed by regularisation methods and with the use of cross-validation (Biganzoli, Boracchi, Mariani, & Marubini, 1998; Lisboa, Wong, Harris, & Swindell, 2003; Ripley, 1996; Ripley & Ripley, 2001). The second key issue is transparency, i.e. explaining what influences the network predictions and how to resolve outcome predictions in terms of readily understood clinical statements. This is partly addressed by rule-extraction algorithms.

Abbreviations: CTs, clinical trials; RCTs, randomised-controlled trials; AI, artificial intelligence; ANNs, artificial neural networks; SOMs, self-organised maps; PCA, principle component analysis; MLC, maximum likelihood classifiers; SVM, support vector machines; FLD, fisher linear discriminators; LR, linear regression; GA, genetic algorithms; CART, classification and regression tree; MVDA, multivariate discriminate analysis.

* Corresponding author. Tel.: +44 151 706 4214; fax: +44 151 706 5803.
E-mail address: afgt@liv.ac.uk (A.F.G. Taktak).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.10.007

Notwithstanding these concerns, an interesting feature of neural network decision support in medicine is the routine clinical use of a range of systems, from the commercial-C.Net (Nabney, Evans, Tenner, & Gamlyn, 2001) and BioSleep (Tarassenko, McGrogan, & Braithwaite, 2002)—to research prototypes (Lisboa, Ifeachor, & Szczepaniak, 2000; Taktak, Fisher, & Damato 2004) without listing in PubMed of supportive clinical trials. The situation is not specific to neural networks, but extends particularly to web-based decision support tools such as www.adjuvantonline.com, marking a departure from algorithms for clinical routine assessments, e.g. the Glasgow Coma Score for severity of illness in critical care and Nottingham Prognostic Index for breast cancer, both of which have undergone rigorous multi-centre clinical trials evidenced in the literature, if not altogether without controversy.

The use of unstructured approaches to clinical evaluation of new medical research is a trend, which has proved hard to change. Already in 1994 a paper entitled 'the scandal of poor medical research' (Altman, 1994) highlighted the need to proper study design bordering on the unethical typically through the application of such bad scientific methodology as to be sometimes called 'torturing the data' until they confess to the desired result (Mills, 1993).

Therefore, it is important to define and keep to a staged framework to design a sequence of studies each with a clear-cut purpose, ranging from the exploratory to the definitive, where the chief aim of each step in this chain is to support the next developmental step until a power calculation is possible which will determine the sample size, along with clinical protocol and study design for a multi-centre randomised clinical trial. Such a framework has been published (Campbell et al., 2000) and adapted for the development of intelligent decision support in an earlier review (Lisboa, 2002). This review will note the current trends in the studies that reach journals in the medical or medically related science literature, highlight points of good and poor practice, and draw conclusions for study design to improve the likelihood of studies being appropriately followed-up in the future.

2. Literature search

A systematic literature search was conducted using Pubmed for entries during the period 1994–2003 with the keywords 'neural networks'. The search was limited to clinical trials and randomised controlled trials (RCTs). Results of the search are summarised in Fig. 1. The search was repeated using the keywords (neural networks) and (cancer) from 1994 to the current date. There were 396 hits in total with only 27 either CTs or RCTs and the abstracts of the resulting hits were analysed. The effectiveness of this technique has been shown to have 50–60% sensitivity (Gant et al., 2001).

Some trials showed clear added benefit in using ANNs whilst others were only able to show that they performed as well as traditional methods. A third group showed that there are advantages and disadvantages in using ANNs. The trials were therefore classified into two categories:

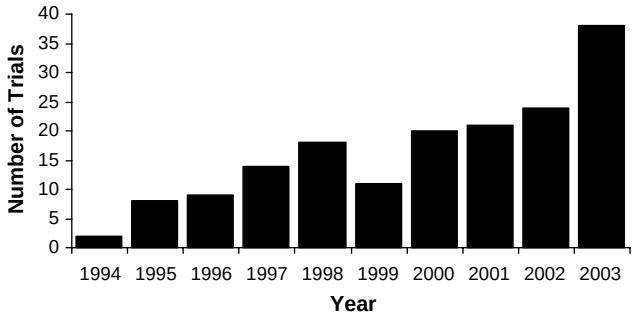

Fig. 1. Number of clinical trials involving the use of ANNs in clinical diagnosis in the last decade.

(a) Those that showed an added benefit containing the first group.
(b) Those that did not contain the last two groups.

None of the trials examined showed conclusive decrease in benefit in using these techniques. The number of subjects and the main findings in each trial were also noted as a measure of the statistical power of the study and was plotted in a funnel graph according to their category, shown in Fig. 2. The values in the abscissa represent the total number of patients/samples included in the study (subjects and controls).

The number of published papers was further compared with the incidence of cancer (Parkin, Whelan, Ferlay, Raymond, & Young, 1997). The plot in Fig. 3 shows a preponderance of publications on cancers of the prostate and cervix, arguably because there is considerable potential for patient benefit from well-tailored therapy, compared to, for example, cancer of the lung. There is also a higher than expected proportion of publications on rare diseases, arguably exploiting a need for greater decision support in areas where clinical expertise is scarce.

3. Review of papers related to cancer listed in Pubmed

The majority of clinical trial studies benchmarked the ANNs performance against traditional screening methods. In prostate cancer, this involves the use of prostate specific antigen (PSA) serum marker, digital rectal examination, Gleason sum, age and race (Gamito, Stone, Batuello, &

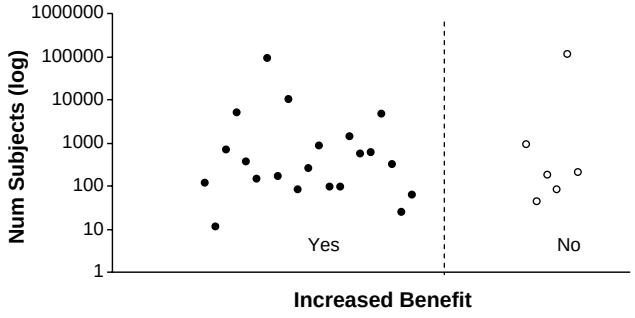

Fig. 2. Trials that showed an increase in benefit using ANNs in cancer (black circles) and those that did not (white circles).

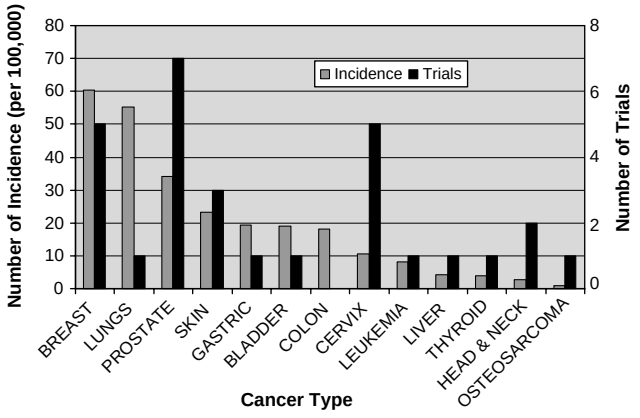

Fig. 3. Number of incidence of different types of cancer in Europe on the left axis and number of clinical trials for each type on the right axis.

Crawford, 2000; Remzi et al., 2003; Stephan et al., 2003; Tewari et al., 2001). Some studies have compared ANNs with statistical methods (Chan et al., 2003; Finne et al., 2000; Matsui et al., 2002; Remzi et al., 2003). Remzi demonstrated that ANNs are more accurate than multivariate logistic regression (LR) using ROC analysis and therefore reduced the number of unnecessary repeat biopsies. They went on to conclude that their system would allow individual counselling of patients with an initial negative biopsy. Finne on the other hand showed that ANNs and LR are both accurate than PSA alone and also reached the conclusion that they reduced the number of unnecessary repeat biopsies.

Cervical cancer applications concentrated mainly on evaluating the benefits of the widely known PAPNET system (Doornewaard et al., 1999; Kok & Boon, 1996; Mango & Valente, 1998; Nieminen, Hakama, Viikki, Tarkkanen, & Anttila, 2003; Sherman et al., 1997), one of very few ANNs systems to gain FDA approval for clinical use. The system uses ANNs to extract abnormal cell appearance from vaginal smear slides and describe them in histological terms (Boon & Kok, 2001). The alternative more conventional way is to re-screen the slides under the microscope. Mango and Valente have shown that the PAPNET system uncovered a higher proportion of false negatives than conventional microscopic re-screening as confirmed by cytologists. Sherman looked at the results of PAPNET in 200 specific cases where initial screening was inconclusive and compared them with conventional microscopy, DNA analysis and biopsy. The study showed that for these cases, PAPNET would have reduced unnecessary biopsies but at the expense of increasing false positives.

Parekattil, Fisher, and Kogan (2003) showed in a clinical trial on bladder cancer that their ANNs model was more accurate in identifying patients who required cystoscopy thereby providing possible savings. Whilst the majority of ANNs algorithms used in the trials concentrated on multi-layer perceptrons (MLP), one study used a hybrid system combining non-supervised Kohonen self-organising map (SOM) with MLP (Glass & Reddick, 1998) to study the response of paediatric osteosarcoma to chemotherapy using MRI images. The technique showed a high correlation with histopathologic analysis. One study used ANNs as a validation method for another technique, namely electrical impedance spectroscopy to separate basal cell carcinoma (BCC) from benign skin lesion (Dua, Beetner, Stoecker, & Wunsch, 2004).

There are several aspects of good practice, not least the proportion of studies involving more than 200 subjects, of which there are 14. However, only five publications carried out regularisation on a principled basis, Fujikawa et al. (2003) with the Bayesian evidence approximation, Glass et al. (1998) using a hybrid combination of self-organising map (SOM) and a multi-layered perceptron (MLP), Gletsos et al. (2003) using genetic algorithms (GA), Chan et al. (2003) and Lin et al. (2004) using a support vector machine (SVM), the remaining 22 studies apparently relying on vanilla MLPs.

Moreover, performance assessment consisted of estimating misclassification rates from a single train/test split in 20 out of the 27 papers reviewed, the exceptions being three RCTs (Gamito et al., 2000; Matsui et al., 2002; Remzi et al. 2003) which used separate train/test/validation datasets, two RCTs (Chan et al., 2003; Finne et al., 2000) and one CT (Lin et al., 2004) that applied leave-one-out cross validation, and one RCT (Bryce et al., 1998) that applied round-robin cross-validation. Only one trial using Papnet (Nieminen et al. 2003) was a prospective study.

While several papers express classification performance as the area under ROC (AUROC), also quoting values of sensitivity, specificity and positive predictive value for the threshold of choice, very few applied rigorous tests to compare their method with benchmark systems. Remzi et al. (2003) compared their diagnostic system with the benchmark logistic regression model and conventional PSA tests using the McNemar test modified by Bonferoni-Holm. The use of AUROC to claim a performance advantage is also invalidated by the absence of confidence intervals for the area values with the exception of Bryce et al. (1998).

A complete list of CTs and RCTs with the clinical application, number of subjects and methods used is presented in Tables 1 and 2 respectively. As explained earlier, some studies showed a clear added benefit in the use of ANNs techniques in cancer diagnosis whereas others did not. Fig. 2 separates these two groups and shows them plotted against the number of subjects in the trial as a method of assessing the statistical power of such trial.

Overall, the publications reviewed were favourable to the neural network approaches, although two of the most proficient studies, both about prostate cancer, drew conflicting conclusions results from very similar empirical results (Matsui et al., 2002; Remzi et al., 2003). However, there is clearly some way to go before establishing the case for a performance advantage for neural networks over conventional statistical methods in the diagnosis of complex data, a finding that is supported by reviews of prostate cancer noting that:

> "…priority areas to be addressed are medical record quality, the need for proper evaluation of repeatability through

Table 1
Summary information from clinical trials (CT) involving the use of ANNs in cancer

Reference	Application	Subjects	Methods
Diagnosis			
Dua et al. (2004)	Skin	34	ANNs
Mango and Valente (1998)	Cervical	10,000	PAPNET
Sherman et al. (1997)	Cervical	200	PAPNET
Stephan et al. (2003)	Prostate	94	ANNs
Tewari et al. (2001)	Prostate	1400	ANNs
Tomatis et al. (2003)	Skin	534	MVDA, ANNs
Vomweg et al. (2003)	Breast	604	ANNs
Prognosis			
Image analysis			
Coppini et al. (2003)	Lungs	312	ANNs
Lin et al. (2004)	Cervical	59	SVM, PCA
Stefaniak, Cholewinski, & Tarkowska (2003)	Head and Neck	25	ANNs

Table 2
Summary list of randomised controlled-trials (RCT) involving the use of ANNs in cancer

Reference	Application	Subjects	Methods
Diagnosis			
Chan et al. (2003)	Prostate	11	MLC, SVM, FLD
Doornewaard et al. (1999)	Cervical	898	PAPNET
Finne et al. (2000)	Prostate	656	ANNs, LR
Gamito et al. (2000)	Prostate	5099	ANNs
Kok and Boon (1996)	Cervical	91,294	PAPNET
Matsui et al. (2002)	Prostate	178	ANNs, LR
Naguib et al. (1996)	Breast	81	ANNs
Nieminen et al. (2003)	Cervical	108,686	PAPNET
Parekattil et al. (2003)	Bladder	253	ANNs
Simpson et al. (1995)	Breast	91	FLD, ANNs
Prognosis			
Bryce et al. (1998)	Head and neck	116	ANNs
Kothari, Cualing, and Balachander (1996)	Leukemia	170	ANNs
Remzi et al. (2003)	Prostate	820	ANNs, GA
Image analysis			
Gerger, Pompl, and Smolle (2003)	Skin	369	CART, Machine learning
Glass and Reddick (1998)	Paediatric osteosarcoma	43	SOM, ANNs
Gletsos et al. (2003)	Liver	147	ANNs, GA
Ng, Ung, Ng, and Sim (2001)	Breast	82	ANNs

multicentre studies quoting results in terms of the ROC framework as well as resistance to change in working practice especially from older clinicians…. ANNs do not prove always better as to replace standard statistical analysis as the method of choice in interpreting medical data. In particular, likelihood odds ratios calculated separately for each explanatory variable are considered by clinicians to be easily understood and useful, a way of opening the black-box" (Anagnostou, Remzi, Lykourinas, & Djavan, 2003).

It is widely acknowledged that neural network modelling requires large amounts of data, moreover they "… have not fulfilled the expectation of some proponents that it would eclipse more conventional statistical techniques." (Lynch et al., 2001), moreover "…several issues associated with neural network derivation demand that developers apply rigorous engineering practices in their studies." (Rodvold, McLeod, Brandt, Snow, & Murphy, 2001).

While parallel studies have identified neural network methods among the most prevalent non-traditional methodologies for data analysis (Chau, 2001), in realising their potential application needs to overcome obstacles including the need for expanded databases and the need to establish multidisciplinary teams (Dayhoff & DeLeo, 2001) and lack of appropriate gold standards (Zhang, Huang, & Roy, 2002)

4. Implications for study design

It is well documented that hundreds of papers are published in the medical literature, at a vast mean cost per published paper, yet few results find their way into improving healthcare practices in routine clinical use. There are reasons for this, partly the unavoidable result that not all interesting new methods turn out to fulfil their early promise. However, more often than not it is methodological shortcomings that mortally damage the future worth of the paper. Some of the reasons for this are (Altman & Royston, 2000; Wyatt & Altman, 1995):

- Study aims with little immediate clinical relevance.
- Model structures that lack clinical credibility.
- Lack of clear purpose for the study, in particular distinguishing between,
  - exploratory studies that aim to generate insights into the data and to optimise model complexity, typically with retrospective studies which need to have proper evaluation strategies, and should explain how the method will integrate into clinical processes, and
  - pragmatic studies aiming to establish the utility of a predictive model, typically prospective studies, whose results should include the definition of a clinical need, an indication of what performance is needed in order to achieve clinical usefulness.
- Overoptimistic assessment of predictive performance.
- Poor model selection procedures when many variables are involved.
- Insufficient estimation due to small sample sizes, which should match a ratio of events of interest (e.g. diagnosed

cancer cases) *per* variable of interest (e.g. *possible* input variables to choose from) of the order of 10.
- Poor evaluation and benchmarking, typically
  - quoting the 'best model' with a train/test validation introduces a bias causing an underestimate of the error rate as the test data are used for model selection,
  - lack of proper statistical performance measures to compensate for the effect of prevalence, e.g. the ROC framework,
  - ad hoc comparisons between neural network performance against a benchmark, when statistical methodologies need to be used if comparative performance claims are to be made (Ripley, 1996).

All of these shortcomings are apparent in the medical statistics literature in general, but especially so for papers involving neural networks (Schwartzer, Vach, & Schumacher, 2000). These concerns can be addressed by consideration of the following practices:

- Bias in the estimation of error rates can be avoided by optimising the neural network model with cross-validation methods instead of train/test splits, employing a separate dataset for independent estimation of the error, which should ideally be a distinct cohort either later in time (temporal validation) or from different clinical centres (external validation).
- Efficient performance estimates do sufficient sample size, whose effect should be quantified by means of 95% confidence intervals for a full range of ROC values including sensitivity, specificity and positive predictive value for the classified of choice, as well as the AUROC for the model.
- Implausible decision functions result from overfitting, which needs to be controlled through a combination of:
  - appropriate sample size in terms of EPV,
  - use of a principled regularisation scheme,
  - a method to explain the network's response, e.g. sensitivity analysis, log-odds ratio, or rule extraction.
- Information on network complexity has to be quoted, including the number of nodes per layer and hence the free parameters (weights) as a ratio of the number of events of interest—where the evidence approximation is use (Bishop, 1995) then the approximate number of free degrees of freedom actually utilised by the network may also be cited.
- Appropriate benchmarking should include gold standard clinical assessment procedures, as well as credible alternative statistical and machine learning models, e.g. logistic regression, nearest neighbours, CART, etc.
- Insufficient comparisons with benchmark performance undermine the conclusions from the study, therefore McNemar or similar tests are required (Ripley, 1996) and for comparisons between particular classifiers 95% confidence intervals may be obtained about individual ROC points (Tilbury, Van Eetvelt, Garibaldi, Curnow, & Ifeachor, 2000).

These elements of good practice go a long way towards maximising the benefit of clearly defined studies set within a staged framework for the development of decision support systems in medicine, a good example of which is the continuum of evidence outlined in Table 3. This framework enables a clear purpose and methodology to be defined for different types of study at each stage in the ladder, enabling the conclusions of one study to feed into the next.

5. Ethical and legal issues

A final consideration with particular implications for the evaluation of biomedical decision support systems concerns the legal and ethical foundation to judge whether the 'duty of care' has been breached. The principles involved hark back to the 'Bolam test' which refers to the skill of an ordinary competent practitioner. This test offers considerable latitude in the exercise of clinical discretion, a leniency founded on confidence in the doctor's training (Gant et al., 2001). Similar considerations apply also to regulatory requirements for evidence of repeatability, reliability and performance.

The first question is: are such systems considered as a medical device and therefore subject to the medical device directive (MDD) and CE marking? Article 1 in the Directive defines a medical device as "any instrument, apparatus, *appliance, material or other article* whether used alone or in combination, including the software necessary for its proper application intended by the manufacturer to be used for human beings for the purpose of: diagnostic, prevention, monitoring, treatment or alleviation of disease,…" This suggests that software whether standalone or part of a device can be a medical device.

Table 3
MHRA view on the use of software in medical applications (MEDDEV 2.1/1)

Software that constitutes a medical device	Software which is not a medical device
Control or influence the functioning of a medical device	Administrative handling
Use for/by patients to diagnose or treat a physical or mental condition or disease	Education software
Analysis of patient data generated by a medical device with a view to diagnosis and monitoring	Maintenance of medical devices or components of medical devices
	Design and manufacturing processes of the medical device. (e.g. compilers, CM systems, MRP, production control, inventory control, SPC, etc.)
	'Operating system', support or system software

The view of the UK regulatory body, the Medicines and Healthcare Products Regulatory Agency (MHRA) on software is summarized in Table 3 (MEDDEV 2.1/1). Taking this into account therefore, the type of ANNs and other AI systems covered in this review would probably fall under the definition of a medical device. Suppliers of medical devices are under legal obligation to demonstrate that the device meets the Essential Requirements detailed in Annex I in the MDD. The most reliable way of ensuring that a device complies with the Essential Requirements is to ensure its compliance with the appropriate harmonized standards such as the IEC 60601-1-4 in the case of software.

Unfortunately none of these or any other current European standards make a special case for the incorporation of AI in software so it is difficult to know how to meet the Essential Requirements for such systems. However, the Food and Drug Agency in the USA have issued a guidance document for software in medical devices, which is based on the European IEC 60601-1-4 standards (FDA, 1998). The document has a section on expert systems and ANNs software, which provides useful tips for manufacturers and assessors. Some of the points highlighted by this document are that ANNs can behave in a non-deterministic manner. The designer should therefore justify and explain the choices made for the artificial neural network model, topology, and training sets, as well as explain and justify the data set class that the ANNs is intended to analyze or process. The designer should also describe how overfitting was avoided and should demonstrate how the relevant features were extracted (such as a specific pattern to be detected) and not a peculiarity contained only in the training set. The document sets out a requirement for additional data sets to be processed through the network to ensure that the performance remains as expected.

What does this mean for the use of artificial neural network models in medicine? An immediate implication is to require rigorous evaluation of their:

- Accuracy in regression or classification
- Repeatability or generality of performance
- Transparency in relation to clinical knowledge, meeting the requirements from the doctrine of 'learned intermediaries'

Further evaluation is needed in relation to the specific rôle of decision support, which can be generally categorised as:

- Guiding patient management by means of
  - Inferences about diagnosis, prognosis or treatment effects
  - Summaries of complex low-level data, e.g. through visualisation
- Filtering of similar cases to the individual patient at hand
- Alerting for rare events, e.g. high-risk cases
- Auditing compliance with, and more importantly, consistency in the use of clinical discretion within clinical guidelines.

In this context, the evaluation of models to guide patient management, in particular, must include an additional and demanding measure to ensure that the model inferences accurately fulfil the role probabilities namely, calibration which is the equivalent of a correlation plot of prediction vs. outcomes in regression, which becomes a plot of the prevalence of outcome against the predictions from the model (Lisboa, Vellido, & Wong, 2000). This is especially important as decision support needs to move away from 'oracle systems' and turn towards informing, rather than advising, the clinician, e.g. by inferring risk of disease and presenting it as a colour-coded bar length to indicate that out of 100 similar patients a certain proportion would be expected to have the condition.

With regard to the evaluation of decision systems, considerations about regulatory and legal aspects have implications that extend to the very purpose of systems intended to help with medical decision making, leading away from the treatment of diagnostic support as a mirror for the statistical analysis relevant to therapeutic studies. In recognising the decision making role of the clinician, computerised decision support systems serve not to instruct on a decision on a predicted outcome, but to modulate the clinician's own decision by adding new evidence through associative recall from historical data. A practical framework to put this into practice is a Bayesian approach where the likelihood ratio of the diagnostic test factors into the clinician's pre-test likelihood of a diagnosis, returning a posterior probability that the clinician may test in multiple scenarios by considering the effect over a range of plausible choices of pre-test likelihoods derived from the available clinical signs.

6. Conclusions

A review of PubMed listed publications involving clinical trials of neural network systems identified trends in areas of clinical promise, specifically in the diagnosis, prognosis and therapeutic guidance for cancer, but also the need for more extensive application of rigorous methodologies. This has implications for study design, to address some of the more common pitfalls of empirical models for medical diagnosis, particularly those relying on generic non-linear function approximations, which includes artificial neural networks.

Further considerations regarding evaluation of systems for clinical decision support led to a categorisation of possible functional roles from auditing, through personalised information about 'cases like yours' and filtering of rare conditions, to clinical guidance. Issues of regulatory compliance, legal recourse and clinical acceptance all point towards an 'added-value' framework whereby the diagnostic test forms an extension of standard laboratory tests adding information to modulate the clinician's own scenarios about likelihood of a particular diagnosis, which can be practically implemented within a likelihood ratio framework.

Acknowledgements

This work is supported by the BIOPATTERN EU Network of Excellence. EU Contract 508803.

References

Altman, D. G. (1994). Editorial: The scandal of poor medical research. *British Medical Journal*, *308*, 283–284.

Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, *19*, 453–473.

Anagnostou, T., Remzi, M., Lykourinas, M., & Djavan, B. (2003). Artificial neural networks for decision-making in urologic oncology. *European Urology*, *43*(6), 596–603.

Biganzoli, E., Boracchi, P., Mariani, L., & Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine*, *17*(10), 1169–1186.

Bishop, C. M. (1995). *Neural network for pattern recognition*. Oxford: Clarendon Press.

Boon, M. E., & Kok, L. P. (2001). Using artificial neural networks to screen cervical smears: How new technology enhances health care. In V. Grant, & R. Dybowski (Eds.), *Clinical applications of artificial neural networks* (pp. 81–89). Cambridge: Cambridge University Press.

Bryce, T.J., Dewhirst, M. W., Floyd, C.E., Jr., Hars, V., & Brizel, D.M., 1998, Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck. *International Journal of Radiation Oncology. Biological Physics, 41*(2), 344–345.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., Sandercock, P., Spiegelhalter, D., et al. (2000). Framework for design and evaluation of complex interventions to improve health. *British Medical Journal*, *321*, 694–696, www.mrc.ac.uk/complex_packages.html.

Chan, I., Wells, W., Mulkern, R. V., Haker, S., Zhang, J., Zou, K. H., et al. (2003). Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier. *Medical Physics*, *30*(9), 2390–2398.

Chau, T. (2001). A review of analytical techniques for gait data. Part 2: Neural network and wavelet methods. *Gait Posture*, *13*(2), 102–120.

Coppini, G., Diciotti, S., Falchini, M., Villari, N., & Valli, G., 2003, Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms. *IEEE Transactions on Information Technology in Biomedicine, 7*(4), 344–357.

Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural networks: Opening the black box. *Cancer*, *91*(8 Suppl.), 1615–1635.

Doornewaard, H., van der Schouw, Y. T., van der, G. Y., Bos, A. B., Habbema, J. D., & van den Tweel, J. G. (1999). The diagnostic value of computer-assisted primary cervical smear screening: A longitudinal cohort study. *Modern Pathology*, *12*(11), 995–1000.

Dua, R., Beetner, D. G., Stoecker, W. V., & Wunsch, D. C. (2004). Detection of basal cell carcinoma using electrical impedance and neural networks. *IEEE Transactions on Biomedical Engineering*, *51*(1), 66–71.

FDA. (1998). *Guidance for FDA reviewers and industry: Guidance for the content of premarket submissions for software contained in medical devices*. US Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Office of Device Evaluation.

Finne, P., Finne, R., Auvinen, A., Juusela, H., Aro, J., Maattanen, L., et al. (2000). Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network. *Urology*, *56*(3), 418–422.

Fujikawa, K., Matsui, Y., Kobayashi, T., Miura, K., Oka, H., Fukuzawa, S., et al., 2003, Predicting disease outcome of non-invasive transitional cell carcinoma of the urinary bladder using an artificial neural network model: Results of patient follow-up for 15 years or longer. *International Journal of Urology, 10*(3), 149–152.

Gamito, E. J., Stone, N. N., Batuello, J. T., & Crawford, E. D. (2000). Use of artificial neural networks in the clinical staging of prostate cancer: Implications for prostate brachytherapy. *Techniques in Urology*, *6*(2), 60–63.

Gant, V., Rodway, S., & Wyatt, J. (2001). Artificial neural networks: Practical considerations for clinical applications. In V. Gant, & R. Dybowski (Eds.), *Clinical applications of artificial neural networks* (pp. 329–356). Cambridge: Cambridge University Press.

Gerger, A., Pompl, R., & Smolle, J. (2003). Automated epiluminescence microscopy—Tissue counter analysis using CART and 1-NN in the diagnosis of melanoma. *Skin Research and Technology*, *9*(2), 105–110.

Glass, J. O., & Reddick, W. E. (1998). Hybrid artificial neural network segmentation and classification of dynamic contrast-enhanced MR imaging (DEMRI) of osteosarcoma. *Magnetic Resonance Imaging*, *16*(9), 1075–1083.

Gletsos, M., Mougiakakou, S. G., Matsopoulos, G. K., Nikita, K. S., Nikita, A. S., & Kelekis, D. (2003). A computer-aided diagnostic system to characterize CT focal liver lesions: Design and optimization of a neural network classifier. *IEEE Transaction on Information Technology in Biomedicine*, *7*(3), 153–162.

Kok, M. R., & Boon, M. E. (1996). Consequences of neural network technology for cervical screening: Increase in diagnostic consistency and positive scores. *Cancer*, *78*(1), 112–117.

Kothari, R., Cualing, H., & Balachander, T. (1996). Neural network analysis of flow cytometry immunophenotype data. *IEEE Transactions on Biomedical Engineering*, *43*(8), 803–810.

Lin, W., Yuan, X., Yuen, P., Wei, W. I., Sham, J., Shi, P., et al. (2004). Classification of in vivo autofluorescence spectra using support vector machines. *Journal of Biomedical Optics*, *9*(1), 180–186.

Lisboa, P. J. (2002). A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks*, *15*(1), 11–39.

Lisboa, P. J., Wong, H., Harris, P., & Swindell, R. (2003). A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine*, *28*(1), 1–25.

Lisboa, P. J. G., Ifeachor, E. C., & Szczepaniak, P. S. (Eds.). (2000). *Artificial neural networks in biomedicine*. London: Springer.

Lisboa, P. J. G., Vellido, A., & Wong, H. (2000). Bias reduction in skewed binary classification with Bayesian neural networks. *Neural Networks*, *13*, 407–410.

Lynch, J. H., Batuello, J. T., Crawford, D. E., Gomella, L. G., Kaufman, J., Petrylak, D. P., et al. (2001). Therapeutic strategies for localized prostate cancer. *Revista de Urologia*, *3*(2), S39–S48.

Mango, L. J., & Valente, P. T. (1998). Neural-network-assisted analysis and microscopic rescreening in presumed negative cervical cytologic smears: A comparison. *Acta Cytologica*, *42*(1), 227–232.

Matsui, Y., Egawa, S., Tsukayama, C., Terai, A., Kuwao, S., Baba, S., et al. (2002). Artificial neural network analysis for predicting pathological stage of clinically localized prostate cancer in the Japanese population. *Japanese Journal of Clinical Oncology*, *32*(12), 530–535.

MEDDEV 2.1/1, http://europa.eu.int/comm/enterprise/medical_devices/meddev/index.htm.

Mills, J. L. (1993). Data torturing. *The New England Journal of Medicine*, *329*, 116–119.

Nabney, I. T., Evans, D. J., Tenner, J., & Gamlyn, L. (2001). Benchmarking beat classification algorithms. *Computers in Cardiology* , 529–532.

Naguib, R. N., Adams, A. E., Horne, C. H., Angus, B., Sherbet, G. V., & Lennard, T. W. (1996). The detection of nodal metastasis in breast cancer using neural network techniques. *Physiological Measurement*, *17*(4), 297–303.

Ng, E. Y., Ung, L. N., Ng, F. C., & Sim, L. S. (2001). Statistical analysis of healthy and malignant breast thermography. *Journal of Medical Engineering and Technology*, *25*(6), 253–263.

Nieminen, P., Hakama, M., Viikki, M., Tarkkanen, J., & Anttila, A. (2003). Prospective and randomised public-health trial on neural network-assisted screening for cervical cancer in Finland: Results of the first year. *International Journal of Cancer*, *103*(3), 422–426.

Parekattil, S. J., Fisher, H. A., & Kogan, B. A. (2003). Neural network using combined urine nuclear matrix protein-22, monocyte chemoattractant protein-1 and urinary intercellular adhesion molecule-1 to detect bladder cancer. *The Journal of Urology*, *169*(3), 917–920.

Parkin, D. M., Whelan, S. L., Ferlay, J., Raymond, L., & Young, J. (1997). *Cancer incidence in five continents*, (Vol. VII(143)). Lyon: IARC Scientific Publication.

Remzi, M., Anagnostou, T., Ravery, V., Zlotta, A., Stephan, C., Marberger, M., et al. (2003). An artificial neural network to predict the outcome of repeat prostate biopsies. *Urology*, *62*(3), 456–460.

Ripley, B. D. (1996). *Pattern recognision and neural networks*. Cambridge: Cambridge University Press.

Ripley, B. D., & Ripley, R. M. (2001). Neural networks as statistical methods in survival analysis. In V. Gant, & R. Dybowski (Eds.), *Clinical applications of artificial neural networks* (pp. 237–255). Cambridge: Cambridge University Press.

Rodvold, D. M., McLeod, D. G., Brandt, J. M., Snow, P. B., & Murphy, G. P. (2001). Introduction to artificial neural networks for physicians: Taking the lid off the black box. *Prostate*, *46*(1), 39–44.

Schwartzer, G., Vach, W., & Schumacher, M. (2000). On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Statistics in Medicine*, *19*, 541–551.

Sherman, M. E., Schiffman, M. H., Mango, L. J., Kelly, D., Acosta, D., Cason, Z., et al. (1997). Evaluation of PAPNET testing as an ancillary tool to clarify the status of the 'atypical' cervical smear. *Modern Pathology*, *10*(6), 564–571.

Simpson, H. W., McArdle, C., Pauson, A. W., Hume, P., Turkes, A., & Griffiths, K. (1995). A non-invasive test for the pre-cancerous breast. *European Journal of Cancer*, *31A*(11), 1768–1772.

Stefaniak, B., Cholewinski, W., & Tarkowska, A., 2003, Application of artificial neural network algorithm to detection of parathyroid adenoma. *Nuclear Medicine Review Central East Europe, 6*(2), 111–117.

Stephan, C., Vogel, B., Cammann, H., Lein, M., Klevecka, V., Sinha, P., et al. (2003). An artificial neural network as a tool in risk evaluation of prostate cancer. Indication for biopsy with the PSA range of 2–20 microg/l. *Der Urologe. Ausg. A.*, *42*(9), 1221–1229.

Taktak, A. F. G., Fisher, A. C., & Damato, B. (2004). Modelling survival after treatment of intraocular melanoma using artificial neural networks and Bayes theorem. *Physics in Medicine and Biology*, *49*(1), 87–98.

Tarassenko, L., McGrogan, N., & Braithwaite, E. Sleep and its measurement: traditional and contemporary approaches. Proceedings of the Royal Society of Medicine (2002). London, 9 may.

Tewari, A., Issa, M., El Galley, R., Stricker, H., Peabody, J., Pow-Sang, J., et al. (2001). Genetic adaptive neural network to predict biochemical failure after radical prostatectomy: A multi-institutional study. *Molecular Urology*, *5*(4), 163–169.

Tilbury, J. B., Van Eetvelt, P. W., Garibaldi, J. M., Curnow, J. S., & Ifeachor, E. C. (2000). Receiver operating characteristic analysis for intelligent medical systems—A new approach for finding confidence intervals. *IEEE Transactions on Biomedical Engineering*, *47*(7), 952–963.

Tomatis, S., Bono, A., Bartoli, C., Carrara, M., Lualdi, M., Tragni, G., et al. (2003). Automated melanoma detection: Multispectral imaging and neural network approach for classification. *Medical Physics*, *30*(2), 212–221.

Vomweg, T. W., Buscema, M., Kauczor, H. U., Teifke, A., Intraligi, M., Terzi, S., et al. (2003). Improved artificial neural networks in prediction of malignancy of lesions in contrast-enhanced MR-mammography. *Medical Physics*, *30*(9), 2350–2359.

Wyatt, J. C., & Altman, D. G. (1995). Commentary: Prognostic models; clinically useful or quickly forgotten? *British Medical Journal*, *311*, 1539–1541.

Zhang, X. S., Huang, J. W., & Roy, R. J. (2002). Modeling for neuromonitoring depth of anesthesia. *Critical Reviews in Biomedical Engineering*, *30*(1–3), 131–173.