A review of evidence of health benefit from artificial neural networks in medical intervention

P.J.G. Lisboa

School of Computing and Mathematical Sciences, Liverpool John Moores University
Liverpool L3 3AF, UK
Tel. (0151) 231 2225/3226
Fax. (0151) 207 4594
Email: P.J.Lisboa@livjm.ac.uk

Abstract

The purpose of this review is to assess the evidence of healthcare benefits involving the application of artificial neural networks to the clinical functions of diagnosis, prognosis and survival analysis, in the medical domains of oncology, critical care and cardiovascular medicine. The primary source of publications is PUBMED listings under Randomised Controlled Trials and Clinical Trials. The rôle of neural networks is introduced within the context of advances in medical decision support arising from parallel developments in statistics and artificial intelligence. This is followed by a survey of published Randomised Controlled Trials and Clinical Trials, leading to recommendations for good practice in the design and evaluation of neural networks for use in medical intervention.

MeSH terms: Review; Randomised Controlled Trials; Clinical Trials; decision support systems; prospective studies; diagnosis; prognosis; survival analysis.

List of abbreviations

AAP	Acute abdominal pain decision system
ARD	Automatic relevance determination
ART	Adaptive resonance theory
AUROC	Area under the ROC curve
CART	Classification and regression trees
CDSS	Computer-based decision support systems
CI	Confidence interval
CP	Capsular penetration
TNM	Clinical staging of tumour, lymph nodes affected and distant metastasis
ECG	Electrocardiogram
EEG	Electroencephalogram
EGG	Electrogastrogram
FES	Functional electrical stimulation
FP	False positive rate
fPSA	Free PSA
LDA	Linear discriminants analysis
LogR	Multiple logistic regression
MARS	Multivariate adaptive regression splines
MEG	Magnetoencephalography
MLP	Multi-layer perceptron
MLR	Multiple linear regression
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
PET	Positron emission tomography
PSA	Prostate specific antigen
RCT	Randomized clinical trials
ROC	Receiver operating characteristic
SOM	Self-organised map
TNM	Clinical staging from tumour size, lymph nodes affected and distant metastasis

1. Introduction

1.1. The context of medical decision support methodologies

Artificial intelligence has been proposed as a reasoning tool to support clinical decision-making since the earliest days of computing (Ledley and Lusted, 1959). The theoretical and practical problems encountered have led to important developments in statistics and computer science, but it was only during the last decade of the last century that decision support systems have been routinely used in clinical practice on a significant scale. Some of these involve artificial neural networks. However, before embarking on a specific review of neural network applications, it is important to consider the wider context of existing methods from artificial intelligence and statistical diagnostic aids, whose experience provides a framework to discuss the potential of novel approaches.

Early computer models were intended for general clinical consultation, with the aim of systematising the investigation of a range of possible scenarios. Advanced computing would enable physicians to concentrate where they are most needed, at the patient's bedside, while specialist knowledge would be left to recall systems that can handle the 'encyclopaedic' aspects of medicine (Schwartz, 1970). However, it became apparent that the enormous complexity created by interactions between clinical conditions made a comprehensive scenario analysis intractable. This started a dilemma that is still current, namely the need to specialise the design of decision support systems to closely circumscribed medical problems, when clinicians have no reason to take-up computational tools unless they are useful for almost every patient in a generic category of clinical conditions (Shortliffe, 1993). Worse still, it has long been recognised that a substantial proportion of clinical cases involve rare conditions, e.g. 15% has been reported for acute abdominal pain (Shepherd, 1972), which are discarded by the ruling-in of an assumed set of disease categories, implicitly ruling-out the rest.

Computing research turned towards a formalisation of the process of model-based reasoning, on the heels of studies of problem-solving strategies which indicated that clinical expertise in difficult cases is largely reliant on causal, pathophysiological reasoning (Schwartz *et al*, 1987). This approach was followed in MYCIN (Shortliffe, 1976), which set the foundations for expert systems. Ironically, this methodology has been most successful in prescriptive domains whose scope can be strictly limited, many of which are outside of medicine. Other successful products have developed as laboratory expert systems, which do not intrude directly into the process of clinical care. One such system to diagnose the results of pulmonary function tests (PUFF) is reported to be have used routinely since 1977 and sold to multiple sites (Aikins *et al*, 1983, Coiera, 1997). The clinical potential of expert systems is increasingly being realised in drug prescribing, for instance the project PRODIGY (Prescribing Rationally with Decision Support in General Practice Study) which has undergone extensive multicentre evaluation (Rogers *et al*, 1999). Note that some expert systems, notably MYCIN, use explicit measures of uncertainty and issue prompts for further information, generating an interactive consultation (Shortliffe, 1990). Explanations are provided for the prompts, by listing the information sought to specify the complete set of antecedents for the rules closest to firing.

An alternative approach, closer to classical statistics, was to "construct and evaluate hypotheses by matching a patient's characteristics with stored profiles in a given disease" (Schwartz *et al*, 1987). This approach represents a paradigm shift from structured reasoning modelled on human expertise, which accesses a knowledge base consisting of operational rules, to one based on empirical evidence, relying on representative databases of historical data. A well-known

programme using this approach was INTERNIST-QMR (Miller *et al*, 1982). This programme mirrors hyptothetico-deductive reasoning (Shortliffe,1990) by assigning scores to clinical findings, listed by disease profiles, such that the accumulation of scores corresponded to building the evidence base for a clinical diagnosis. It could handle coexistent diseases, and was remarkably accurate from the start, given that it made strong assumptions about mutual independence of the predictive variables (Shortliffe, 1990). Already, it was noted that it can be difficult to trace the system's recommendations in order to adequately explain them to clinicians, in part due to their "reliance on hard-to-understand numerical scores in dealing with uncertainty" (Kulikowski, 1988).

Other examples of successful Bayesian models are de Dombal's system for the diagnosis of acute abdominal pain (AAP) (de Dombal *et al*, 1972, 1984) and the GLAsgow system for the diagnosis of DYSpepsia (GLADYS, Davies and Owen, 1990). In particular, AAP has been used routinely for several years in a clinical environment (McAdam *et al*, 1990, de Dombal *et al*, 1997). Structured belief networks implemented with the aid of Bayes's theorem have outlasted the 70's and 80's. Where independent Bayes was once used 30 years ago, good approximations of the integrals over distribution spaces are now possible with numerically intensive methods such as Markov Chain Monte Carlo (Neal, 1996). These developments anticipated the current confluence of interest on graphical models (Lunn *et al*, 2000) from sectors of the artificial intelligence, statistical modelling and machine learning communities.

Nevertheless, at the end of the 80's serious deficiencies were recognised in both expert systems and naïve Bayes profile-matching. They were unable to cope with variations in disease presentation, both in terms of the spectrum of findings and their severity. Nor could they deal with the evolution of disease over time, for instance in response to treatment, recognise how one disease may influence the presentation of another, or provide explanations based on patho-physiology (Schwartz *et al*, 1987). Moreover, new knowledge was not readily accommodated without potentially disruptive consequences for the model, in violation of the very principles of adaptation that are characteristic expectations of intelligent systems. These objections still apply today to neural networks, and must be taken into account when considering the rôle of new paradigms for medical decision support.

A third approach to computational reasoning emerged, built around compartmentalised models of anatomy, physiology and pharmacology. Model-based reasoning is expressed as coupled non-linear differential equations or as causal probabilistic networks, for instance to simulate dynamic response to therapy, adapting for particular patients by fitting key parameters to small amounts of data (Andreassen *et al*, 1994). These methods were the focus of decision modelling in the 1980's, taking over from expert systems and empirical data modelling in the hope of resolving the serous concerns about those methods which were outlined earlier, in particular the need to accommodate multiple-disease diagnoses, sequential gathering of evidence over time and the perceived clinical trend to 'delve deeper into the origins of disease' (Lucas, 1997).

A consensus developed that clinical advisory systems must specialise to a narrow clinical domain, which can be addressed at two quite different levels. Practical systems have used shallow associations between observations and a fixed universe of allowed outcomes to formulate a phenomenology of particular classes of disease, applicable in specific clinical settings. An example of this approach is the Acute Physiology and Chronic Health Evaluation (APACHE Knaus *et al*, 1985, 1991, Rowan *et al*, 1994, Gunning and Rowan, 1999) and Simplified Acute Physiology Score (SAPS), which are among widely used statistically derived scoring system for patients receiving critical care. Nevertheless, despite statistical studies supporting the robustness of these algorithms (Lemeshow and Le Gall, 1994), there is controversy surrounding their use in

emergency care wards. This comes, in part, from evidence that use of severity of illness scores can increase mortality predictions when they use physiological variables recorded with the Intensive Care Information Systems (ICIS) (Bosman *et al*, 1998), because automatic charting extracts more abnormal values for most physiological variables than manual charting. In addition, extreme variability in the accuracy of mortality prediction across different patient groups (Moreno *et al*, 1998), has also called into question the validity of using score measures in making treatment decisions for individual patients (Civetta *et al*, 1990, 1992).

The rôle of medical decision support can now be summarised and extended in the context of a review of neural network applications. Adapting from taxonomies of artificial intelligence in medicine (Shortliffe, 1990, Coiera, 1997) there are three general categories of application for neural networks in medical decision support:

1. *Tools for attention focusing*, aiming to detect abnormalities, whether in hospital-based information systems or clinical laboratory systems. Their purpose is to remind the user of diagnoses that might otherwise have been overlooked, or recall rarely occurring disease profiles.

2. *Patient-specific assessments and advice*, started out as tools to support generic consultation by providing diagnostic and prognostic inferences. Given the current reliance on complex medical signals, inference models may be regarded as predictive instruments that sit alongside other medical devices. Early approaches have been dubbed Greek oracle models, on account of the fact that physicians turn to them for advice but 'deferring to the dialogue style and recommendations of the machine' (Shortliffe, 1993). An alternative, more interactive implementation of diagnostic and prognostic indices, is to use them as metrics to access electronic databases, in order to recall the records of historical cases predicted to be the most relevant to inform on a specific patient query.

3. *Interactive tools for critiquing and planning* provide 'black-boards' for scenario analysis, allowing the clinician to gain new insights by testing hypotheses about the patient's condition, and to consider the possible effects of different treatment choices. The utility of this approach can be substantially enhanced by comparing the clinician's inferences with those derived from historical data, by automatically requesting further information about the most informative indicators to resolve discrepancies between the two.

1.2. Previous surveys of medical applications involving neural networks

Early reviews of medical applications of neural networks focused on potential for non-linear data analysis, distributed associative memory function as a was of avoiding difficulties with the acquisition of expert knowledge, noise tolerance due to their inherently parallel architecture, and adaptability to accommodate new manifestations of disease (Reggia, 1993). The thorny issue of performance assessment and clinical evaluation was also starting to be addressed (Ezquerra and Pazos, 1994). Their rôle in the modelling of neurological and psychiatric function was also clearly delineated from that of decision support, where the modelling of empirical data was recognised to be akin to statistical pattern recognition (Reggia, 1993).

A series of articles in an influential medical journal emphasised the prevalence, in clinical practice, of an associative memory function (Cross *et al*, 1995) attuned to the variability of disease presentations and degrees of severity which compound intra- and inter-patient effects. In statistical terms, medical data typically have very low signal-to-noise ratios, as a result of the earlier effects combined with measurement noise, variations in clinical protocols across multiple

centres, and even significant geo-demographical differences between patient populations (Kennedy *et al,* 1997). The high levels of noise typical in medical data cast a long shadow over the appropriateness of finely tuned non-linear models, requiring extensive, and indeed very expensive, performance testing to have confidence in claimed generalisation performance.

The rationale for using neural networks to model the phenomenology of disease was originally to replicate associative memory functions that are known to be important in many familiar activities, even when they involve the exercise of rational judgment in complex tasks (Dreyfus and Dreyfus, 1986). Technologically, however, neural networks remain part of the 'arcane subject of non-linear statistical modelling and inference' (Cross *et al,* 1995) with an intention to make this important and developing field accessible to non-specialist statisticians. It may be argued that sound statistical principles are essential to trust the evidence base built with any data-based methodology, including neural networks (Schwartzer *et al*, 2000). Therefore, these methods are best justified where they provide additional functionality to that of well-established statistical models, which are typically linear-in-the-parameters. This leaves two opportunities. One is to accurately map features of the data that are difficult or expensive to find in a conventional statistical development, typically consisting of complex interactions between particular variables or attributes. The other possibility is to add substantially to the power of exploratory data analysis, for instance by raising hypotheses about unsuspected non-linear components whose explicit modelling may improve the accuracy of standard statistical methods, or by providing direct visualisation of complex high-dimensional data. This review focuses on the current clinical application of neural networks, which is chiefly for non-linear inference, bringing-in their power in exploratory analysis as part of the general discussion.

By 1995 there were close on 1,000 citations of neural networks in the biomedical literature (Baxt, 1995), mostly describing studies on historical data, often small sets on which the predictive accuracy was tested. Certain medical domains were identified to have potential for diagnostic support, at least on the basis of accuracy in single site trials. An example is pre-screening of patients who present to emergency units with acute chest pain, to try to identify those who are most likely to have suffered Acute Myocardial Infarction (AMI). Studies involving 706 patients returned sensitivities and specificities from 80% to 96% (Baxt and White, 1995), compared with a large study of physician's performance that claimed an overall sensitivity and specificity of 88% and 71%, respectively (Goldman *et al*, 1988). Most of these papers used current neural network methodologies, almost invariably the multi-layer perceptron with 'early stopping' to prevent over-fitting. A particularly interesting methodological study concerned the influence of input variables on the response of a network trained to recognise high risk of AMI. This sensitivity analysis uses the bootstrap to correct for small sample bias and shows that the bias is significant (Baxt and White, 1995).

Other promising areas included image analysis and radiography, recovery from surgery and prostate, breast and ovarian cancer, and clinical pharmacology. However, there was little evidence that clinicians would be interested in developing the prototypes further. Given the apparent reluctance of clinicians to readily embrace computer-based decision support (Johnston *et al*, 1994), it is necessary to demonstrate a greater appreciation of the broader clinical needs served by decision support, as well as the results from related studies with alternative methodologies.

The third paper in the same series was concerned with applications in pathology and medical laboratories (Dybowski and Gant, 1995). It is recognised that statistical and neural network methods are ideally suited for cytopathology. Typical diagnostic procedures require he assessment of profiles of cytological features, for example in fine needle aspirates from breast lumps or Paponicolau slides from cervical smears (Koss, 1989). Neural networks were also being

developed to identify malignancies in breast and lung with techniques from clinical chemistry, which combines serum-derived tumour markers, plasma indices and nuclear magnetic resonance spectral measurements.

The Lancet papers pointed out that a contemporary benchmark study, the Statlog project (Comparative Testing of Statistical and Logical Learning, Michie et al, 1994), showed neural network methods to be less accurate than traditional statistics or decision-tree methods for 15 of 22 data sets, three of which consisted of medical data (Wyatt, 1995). This study was intended to relate the performance of 23 discrimination algorithms to characteristics of the data. Some of its results confirmed expectations, for instance, linear and logistic discriminants performed similarly and did better than quadratic discriminants except when there differences in the dispersion of different categories, and that naïve Bayes works best when the indicator variables are approximately independent of one another. Other results were more surprising. The 'locality' of the decision surfaces of the multi-layer perceptron made it similar to a rapid implementation of k-nearest neighbours, which is consistent with a localised associative memory function.

In an attempt to ensure that the results of this large collaborative effort were comparable across the participating centres, and to make a systematic benchmark of a large range of databases practicable, the methods were applied in 'standard' form with relatively little optimisation to each dataset. This protocol afforded only limited control of over-fitting, to which these methods are known to be prone. In relation to medical applications, the authors noted that only a few methods can handle missing data or have the capacity for incremental learning, and machine learning algorithms are the most transparent. Furthermore, the lack of compensation for differences in class prevalence, i.e. heavily skewed data, resulted in the Radial Basis Function (RBF) network on occasions performing worse than the guessing line, while on another example with fewer input variables and a balanced data set it outperformed back-propagation, and it was the best classifier overall for the DNA dataset. So, more detailed benchmarks were needed.

A detailed comparison of multi-layer perceptron, rule induction methods (Quinlan's ID3) and conditional independence Bayes, applied to the diagnosis of acute abdominal pain, gave similar accuracy for all three methods (Schwartz *et al*, 1993). This time, each methodology was carefully optimised for maximum performance on a 'hold-out test 'set comprising 30 cases, from a total design sample of 276. No separate sample was provided to validate the out-of-sample performance estimate. The clinicians' initial diagnostic impressions were compared with the computer-based predictions from the various methods, in the diagnosis of appendicitis against other serious illness and non-specific pain. There were a total of 41 binary indicators, with the ubiquitous missing values coded as separate attributes. The sensitivity and specificity in the differential diagnosis of appendicitis versus the rest indicated that the optimal threshold settings for the MLP and ID3 had similar sensitivity to the doctors but were less specific. The benchmark Bayes classifiers were slightly less sensitive than the MLP, for the same specificity, but all inference methods had much poorer specificity than the doctors. This result is not uncommon, because numerical classifiers tend perform best in the detection of particular clinical conditions, which are well circumscribed, but are poor for characterising diffuse, generic conditions.

A further insight into the diagnostic value of the different procedures may be derived by factorising the ratio of the true positive to false positive ratios, to characterise how the discriminator improves on the guessing line. This defines a boosting factor between the prevalence and the positive predictive value (PPV) of the tests, given by

$$\frac{PPV}{1-PPV} = \frac{Sensitivity}{1-Specificity} * \frac{\Pr(class)}{\Pr(\sim class)}. \qquad (1)$$

Pr(class) and Pr(~class) denote the prevalence within and outside of the diagnostic class of interest. The boosting factors calculated from the sensitivity and specificity values quoted in (Schwartz *et al*, 1993) are 1.1, 1.2, 1.4 and 1.8, for naïve Bayes, ID3, back-propagation and the doctor's initial diagnosis, respectively. The doctors have the clearest gain over the guessing line.

Also of interest is the poor overlap between the models selected by optimising each of the three radically different methodologies. This indicates that, as much as the relevant variables may reflect their inherent symptomatic value, their selection is partly an artefact of the different structures of the discriminant models.

A further detailed benchmark was carried out on a database of 41,021 cardiac patients admitted to 1,081 hospitals in 15 countries (Ennis et al, 1998). The prevalence of death within 30 days of admission was 7%, therefore the class membership is heavily skewed towards survivors. This study followed a detailed design of an effective logistic regression model, with the consequence that some of the predictive variables were composite indices coding for non-linearities in the data, including interaction terms, and there were no missing values as they had been imputed. The first effect militates against the benefit of an additional generic non-linear model.

Generalised additive models, CART and MARS are alternative generic non-linear statistical models also included in the benchmark study. Each model was optimised for the design dataset, which comprised separate subsets for training and parameter tuning, with a further holdout sample of size 13,610 for performance estimation. The multi-layer perceptron architecture comprised an additional layer of direct connections between the input and response nodes, forming a linear model in parallel with the hidden-unit non-linearities. All unregularised multi-layer perceptron networks showed clear evidence of over-fitting, namely a split early on between the mean log-likelihood for the training and tuning, or testing, datasets. This effect is often indicative of a poor signal-to-noise ratio masking any non-linearities that are not be already modelled in the predictor variables. The study concludes that non-linear algorithms have limited applicability in clinical settings, possibly because the signal-to-noise ratio tends to be low. This may result in linear models with interactions asymptotically approximating the Bayesian optimum, with little potential for improvement with generic non-linear models.

So far, the extensive retrospective benchmark studies reviewed have reported no significant differences between the accuracy of neural network models and that of alternative, simpler inference algorithms, including logistic regression. In contrast, the multi-layer perceptron, appears to perform as a sub-optimal generic modelling tool. Nevertheless, the interest on highlighted by recent reviews of applications to critical care (Hanson and Marshall, 2001), surgery (Drew and Monson, 2000, Golub et al, 1998), staging of prostatic cancer (Montie and Wei, 2000), and physical medicine and rehabilitation (Ohno-Machado and Rowland, 1999).

1.3. Scope of the review of publications

In order to make a vast subject manageable, the survey specialises on applications to the key medical domains of oncology, critical care and cardiovascular medicine, as well as the clinical functions of diagnosis, prognosis and survival analysis.

This focus leaves out important application domains, which merit surveys of their own. For instance, mature applications of decision support in medicine often require a combination of several modules operating at different levels, from pattern recognition to supervisory reasoning. This survey provides a snapshot narrowly focused to neural networks, to the exclusion of hybrid methods such as neurofuzzy systems (Szczepaniak et al, 2000). There are also specialist domain areas linked to signal processing that merit focused reviews, such as image processing and time series analysis, including developments in Independent Components Analysis (ICA) for artefact removal and feature extraction (Lisboa *et al,* 2000a). Yet another growing application area closely linked to biomedicine is bioinformatics (Shi et al, 2000).

2. Survey of applications

The purpose of the review of publications is to assess the evidence for improvements in healthcare arising from the involvement of artificial neural networks in medical intervention. It is generally accepted that the most reliable method of determining the effectiveness of a new intervention is to conduct a systematic comparative study with a Randomised Controlled Trial (Jadad and Rennie, 1998). With regard to neural network applications, evidence of impact on healthcare outcomes has also been reported in some observational studies, therefore data collection consisted of a search of publications involving neural networks listed in the PUBMED database (www.ncbi.nlm.nih.gov/PubMed) under Randomised Controlled Trial (RCT) or Clinical Trial (CT). Additional publications of particular interest are also reviewed in the discussion of each medical domain. It is recognised that other, equally meritorious, studies will inevitably have been omitted. Nevertheless, the additional papers provide an indication of significant emerging technical developments.

Medical decision support is different from other medical interventions, such as the introduction of new prescription drugs, because it involves the exercise of clinical judgment, which can substantially affect the study design (Hunt *et al,* 1998, Ohmann *et al,* 1999). For these reasons, a new framework is required to assemble evidence in support of a 'complex intervention', whether it is implemented by means of organisational changes to the delivery of healthcare, through the introduction of new clinical protocols, or by adopting computer-based decision support systems (CDSS).

A flexible framework for design and evaluation of complex medical interventions was recently suggested (Campbell *et al,* 2000). This framework can be used to guide the introduction of new decision support systems, as shown in Fig. 1. Note that each stage feeds directly into a specification for the next with the overall the objective being, not to test a neural network, inference model or any other individual system module, but to assess the outcomes of the intervention as a whole, and demonstrate its merits against a current intervention. This step of comparing with an accepted standard is necessary, not just for clinical acceptance, but also as an essential requirement for certification of medical devices, whose definition includes free-standing software that serves a diagnostic purpose affecting patient care (Lisboa, 2001), through a process that involves formal assessment by an independent certification body.

In reviews of evidence of healthcare benefits arising from medical decision support, it is customary to distinguish between improvements in clinician performance and changes in patient outcome. While clearly linked, the two effects do not always correlate well. For this reason, some surveys filter RCTs and CTs through a system to score the rigour in study design. This filtering was deliberately not carried out in this review, in order to gain a wider perspective of clinical applications of neural networks with a critical assessment of the merits in current practice.

The publications are mapped out by medical domain and clinical function in Tables 1 and 2. The domains of oncology, critical care and cardiology are those where RCTs are prominent, with CTs used for a wide range of medical applications. Most of the studies cited are concerned with diagnostic or prognostic inference, with only one RCT directly addressing survival.

Published reviews of computer-based clinical decision support have concluded that the most promising systems are for drug dosing and preventive care (Hunt *et al*, 1998, Weiner and Pifer, 2000). They serve to alert against adverse effects from prescription drugs, or to promote greater compliance with practice guidelines in health maintenance activities such as vaccinations and mammography. One review paper notes that computer-aided evaluation of mammograms already helps to cut the number of missed lesions by half without increasing the false positive rate (Weiner and Pifer, 2000) though it incurs additional costs in terms of time, training and equipment.

These wide ranging studies do not report any evidence of benefits arising from inference-based decision support despite the routine use, for instance, of several severity of illness scores, notably the Glasgow coma score and APACHE III, the Bayesian decision support systems mentioned. One reason for these omissions is the strict protocol used to select acceptable evidence for the reviews, typically requiring the use of a CDSS in a clinical setting by a healthcare practitioner with assessment by a prospective study against a concurrent control (Hunt *et al*, 1998). While that specification of evidential studies is understandably rigorous, it can be too fine a sieve to fully appreciate the different rôles of statistical modelling and CDSS. An example of a substantial system that is not mentioned in these reviews is AAP. In a letter responding to a survey of computer-assisted diagnosis that similarly omitted this system (Kassirer, 1994) de Dombal pointed out that an eight-centre trial in the United Kingdom involving 16,737 patients with acute abdominal pain showed approximately a 20% improvement in the diagnostic accuracy of the doctors with almost a 50% reduction in the rates of perforation and negative laparotomy (de Dombal, 1994; de Adams *et al*, 1986). A separate study showed comparable results maintained over 12 years (McAdam *et al*, 1990) and this was followed by an international study of 15,000 patients from 64 hospitals in the European Community (De Dombal *et al*, 1993).

Some of these systems are discussed in connection to the review of applications in specific medical domains in the following sections.

2.1. Oncology

2.1.1. Models for inference and visualisation

A list of papers addressing three of the most prevalent cancers is presented in Table 3. The results reported show considerable discriminatory power for quantitative modelling of disease with composite indices comprising indicators that are available in routine tests. This highlights an emerging rôle for quantitative models for decision support in evidence-based medicine.

However, the neural networks tested are generally built around the original implementation of the multi-layer perceptron, with few publication using any form of regularisation of the objective function, for instance weight decay (Khotari *et al*, 1996). One study carried out performance estimation with the bootstrap methodology (Baxt and White, 1995) but internal cross-validation without the use of an additional validation dataset is still very much used to estimate generalisation performance. While this becomes robust for large sample sizes, it is known to return optimistic estimates of misclassification error. These methodological issues are returned to in the discussion of critical design considerations in section 4.

Studies of prostate cancer concern the elimination of false positive Prostate Specific Antigen (PSA) results and the development of prognostic indicators for confirmed cases. PSA concentration is a sensitive but unspecific test for disorders of the prostate. The aim of the study is to increase the specificity without sacrificing the sensitivity, by combining measurements of free PSA and prostate volume, with the outcomes from a digital rectal examination in men aged 55-67 years with total serum PSA concentrations of 4-10 ng/mL (Finne *et al*, 2000). The reduction in false positives was estimated by leave-one-ut cross-validation, at clinically relevant levels of sensitivity 80-99%. The overall conclusion is that either MLP and LogR could reduce the number of biopsies significantly better than using free PSA. Other studies claim that reducing false positives for PSA measurements below 4 ng/mL is also possible with neural networks (Stamey *et al*, 1998).

The other studies address prognostic risk for confirmed diagnoses, and are both contained in a single paper. PSA is again used, this time combined with clinical staging to predict the likelihood of lymph node spread (LNS) (Gamito *et al*, 2000). For capsular penetration (CP), PSA and clinical staging are further combined with PSA velocity and the sum of the Gleason scores for the two most commonly found histological patterns. This is a substantial study showing good practice in the use for validation of an external data set collected from 660 patients, who are additional to the design data from 4,133 patients that were used for training and testing, or tuning, of the classifiers. Once again, the overall conclusion is that accurate objective inference is possible for the likelihood of LNS or CP, although the MLP is not assessed against a benchmark linear classifier. The performance of this staging score is also not compared with alternative scores combining clinical stage, Gleason score and PSA levels, for instance Partin tables (Blute *et al*, 2000).

Cervical cytology studies have generated landmark papers evaluating Neural Network-Assisted (NNA) review of Pap smear slides with the PAPNET Testing System. In 1995 this support system gained Federal Drug Administration (FDA) approval for secondary screening. It is arguably the first concrete medical application of artificial neural networks patented, filed in 1988 and granted in 1990 (Rutenberg, 1990). It also reported the first case-control study where negative smears were re-evaluated using manual microscopic re-screening and PAPNET re-screening (Rosenthal *et al*, 1993). As a retrospective study, it did not merit inclusion in formal reviews of the evidence of benefit from CDSS in medicine.

The PAPNET system (Mango, 1994) identifies for visual inspection the most atypical 64 single cells and 64 clusters among the thousands present in Papanicolaou stained slides taken from cervical smears. This system uses a two-tier architecture with different levels of resolution for initial and final classification. The first tier filters out debris and other unwanted image segments, while the second employing specialist networks for single cell identification and cluster identification. The networks are unregularised multi-layer perceptrons trained by back-error propagation, whose reliability depended entirely on the richness and diversity of samples made available for training and validation. Image features were derived from intensity histograms. In early clinical trials, the agreement between inferences made with PAPNET and a subsequent histological diagnosis was 38% for severe dysplasia, 35% for carcinoma-in-situ and 72% for suspected invasive carcinoma (Boon and Kok, 1995). This compared with 40%, 20% and 62%, respectively, with conventional screening, indicating that the computerised method may significantly improve detection accuracy for the most severe cytological abnormalities. As with AAP twenty years previously, a simple statistical algorithm within a well-structured model has proved to be surprisingly effective and robust. It is clear that the acceptance of PAPNET by laboratory cytologists required more than the performance demonstrated in clinical trials.

Several aspects of the system have contributed to its sustained adoption for routine use. First of all, it addresses an issue where there was an identified need to improvement (Koss, 1989; Sherman and Kelly, 1992). Secondly, the overriding design priority was to substantially improve the sensitivity of diagnostic cell identification, by comparison with trained cytologists, and thus reduce the highest-cost errors caused by false negative tests. The poor specificity, which is invariably traded for improved sensitivity, is managed by referring a set number of images for manual inspection. By rôle specialisation, neural networks thus found a practical purpose for routine use alongside trained cytologists in clinical laboratories. Thirdly, the system promotes user acceptance by integrating seamlessly into the standard testing protocol (Mango, 1997).

Arguably the largest-scale, and most detailed studies of any medical applications of neural networks have been carried out with this system. NNA re-screening of node negative smears produced a statistically significant improvement in yield compared with conventional unassisted re-screening, where the term yield refers to the percentage of re-screened negatives reclassified as abnormal (Mango and Valente, 1998). This conclusion is consistent with an early landmark study (Boon and Kok, 1993) and a large multi-centre study (Koss *et al*, 1997) showing that assisted re-screening could catch cancers that human re-screening missed. A recent CT of inter-observer variability among five cytotechnologists (Sherman *et al*, 1998) showed that it remained high even with NNA screening. Referral of patients with consensus abnormal readings showed a sensitivity of 51% with 31% referrals, which was raised to 95% sensitivity by including also patients with consensus equivocal readings but with 79% referrals.

Since the early trials and subsequent studies reviewing thousands of cases, which demonstrate a potential health benefit arising from an improvement in clinician performance, the capital intensive nature of this technological aid has caused a heated controversy with regard to the cost benefit of using PAPNET (O'Leary *et al*, 1998, Radensky and Mango, 1998). While it is estimated to have a cost per life-year saved that is less than widely used interventions for other conditions such as mammography and PSA (Schechter, 1996), this technology is reported to be more efficient when laboratories read in excess of 50,000 smears per year raising logistical problems for widespread use (Cuzick and Sasieni, 1999).

Nevertheless, the PAPNET system is now being considered for primary screening of cervical cancer. A report of the PRISMATIC team (Prismatic Team, 1999) reports good agreement with conventional primary screening across seven gradings, together with better specificity and faster processing. Another RCT with seven year follow-up in a mass screening programme (Doornewaard *et al*, 1999) also agrees that PAPNET has similar diagnostic value to conventional screening, on the basis of AUROC confidence intervals of 78-82% and 77-81% for conventional and PAPNET screening of dysplasia, respectively. Derivatives of PAPNET are also being applied to the detection bronchogenic carcinoma, from smears of sputum (Koss *et al*, 1996), oesophageal cancer (Koss, 2000) and urithelial carcinoma of the bladder, from bladder washings (Vriesema *et al*, 2000).

In breast cancer, there is interest in surrogate measurement of lymph node status (Naguib *et al*, 1996) and for identification of pre-cancerous breast (Simpson *et al*, 1995). It is clear from these early studies that the quality of data, in particular the occurrence of missing values, is a significant bottleneck for the application of pattern recognition. McGuire *et al*, 1992, carried out a detailed analysis of the prognostic factors influencing predictions of relapse in axillary node-negative patients and propose a framework to use this prognostic information directly to inform treatment decisions. The purpose of this interim study was to further rationalise adjuvant treatment decisions for a patient cohort who are generally at low risk of relapse. Its findings were

that the MLP was sensitive in the detection of the most prevalent group, comprising patients with good-prognosis.

Another study of particular interest addressed the automated cytodiagnosis of fine needle aspirates of the breast from ten cytological features together the patient's age. A prospective data set comprising 322 collected by multiple observers was used to validate a MLP model optimised from a design data set with 692 cases collected retrospectively by a single observer (Cross *et al*, 2000). The sensitivities and positive predictive values at a specificity where there were no false positives, were 79% and 100% when estimated for the MLP from an independent test subset of the design data. However, these values fell to 67% and 91%, respectively, when the model was applied to the external validation data. The 31% drop in sensitivity was considerably worse than the 10% drop observed for logistic regression, from 82% to 72% for the same datasets. This shows that inter-observed variability is a major contributing factor to the effectiveness of automated cyto-diagnosis of fine needle aspirates of the breast, but also highlights the critical rôle for external validation of clinical decision support systems.

In a related study, a new unsupervised neural network architecture was proposed for the visualisation of medical data (Walker *et al*, 1999). The growing cell structure is a dynamic two-dimensional map that grows according to a similarity measure, producing colour maps that highlight the distribution of input variables in relation to the prevalence of disease at different locations in the map. With the use of Bayes' theorem, probability distributions estimated with the Parzen window are converted into class membership probabilities, yielding similar discriminatory performance as logistic regression for breast cytology. Putting aside the evidence of the AUROC, the colour maps of the input features provide a powerful visualisation tool to assess correlations between them and with the externally imposed category label.

Some of the less prevalent cancers have also been the subject of study with neural networks and they are listed in Table 4. A study by Kothari *et al*, 1996, on cell categorization in acute leukemia was the only clinical trial listed to use regularisation of the objective function. Nevertheless, the discrepancy between a negligible training error and a 10% generalisation error, found with various sets of explanatory variables, indicates that either the sample size or choice of design sample makes it insufficiently representative of the test distribution, or the regularisation parameter was underestimated. Bryce *et al*, 1995 and Bugliosi *et al*, 1994 are the only trials directly concerned with prognostic models for survival, as a measure of treatment. Bryce *et al*, 1995 use complete records with two-year follow-up from 116 randomised patients in a Phase III clinical trial comparing hyper-fractionated radiotherapy with or without concurrent systemic treatment. Glass and Reddick, 1998, explore the relatively new modality of contrast enhanced MRI to non-invasively measure a key feature of invasive tumours, the extent of necrosis. Features of interest are identified by a Self-organised map (SOM) and classified by a MLP. Bugliosi *et al*, 1994, applied the MLP but also a holographic model, carrying out automatic variable selection. In particular, 132 further patients were excluded for key missing variables.

2.1.2. Survival analysis

Survival analysis is an important of medical statistics, frequently used to define prognostic indices for mortality of recurrence of a disease, and to study outcome of treatment. While these methods are applied in virtually every medical domain, applications are particularly prevalent in oncology. It has been recognised in the medical literature that neural networks have much to contribute to the modelling of cancer survival (Burke *et al*, 1997, Lundin *et al*, 1999), on the basis of early studies comparing the MLP with the Tumour, Nodes and Metastasis (TNM) clinical

staging system recommended by the World Health Organisation. Initial fied time models were also applied to other areas such as AIDS-related mortality (Ohno-Machado, 1997).

The principal differentiating characteristic of survival modelling compared with conventional class discrimination is the overriding presence of censorship. Censorship occurs when a patient is lost to follow-up without event of interest taking place. In a cancer mortality study, for instance, the event of interest might be death ascribed to breast cancer, with a period of study involving follow-up over five years. Any patient who is lost to the study within five years of recruitment, whether by something as innocuous as changing address or as serious as death from an un-related cause, is considered censored. These inter-current deaths may be particularly difficult to define, as cardiac arrest may be due to systemic damage inflicted during cancer therapy and, as such, the original cancer could legitimately be considered as a contributing factor to the death. Furthermore, all patients reaching the limit of the period of study are deemed censored at the point of maximum follow-up. The term censorship indicates that there is no way of knowing for certain what the outcome for those patients would be had they remained in the study. Treating these data as missing, excluding censored patients from the study or employing *ad hoc* techniques, will incur substantial bias in the estimation of survival, which are readily demonstrated with simulated data (Brown *et al*, 1997).

Several approaches have been proposed to modify the MLP in order to capture the effects of censorship. The motivation for doing so comes from the strict assumptions made in routinely used survival models that are linear in the parameters. Taking as an example the proportional hazards method, also called Cox regression, the assumption is made that for any combination of covariates, the hazard ratio is strictly proportional to that of a selected baseline population. What appears to be a disastrously restrictive assumption is in fact approximately observed in many clinical situations which, combined with the use of a linear prognostic index, makes this model overwhelmingly the most commonly used in large-scale medical statistical studies. As a result, direct MLP extensions of proportional hazards have been proposed (Faraggi *et al*, 1997, Ripley *et al*, 1998) which maintain the separation between the dependence on time and on the patient specific vector of covariates, resulting in non-linear proportional hazards models.

A more efficient representation of time is to include it as a covariate, serving as an input index to condition hazard estimates made by a single output unit. This approach has been proposed by several authors (Ravdin *et* al, 1992, De Laurentiis *et al*, 1994, Liestøl *et al*, 1994), and is thoroughly described by the Partial Logistic Artificial Neural Network (Biganzoli *et al*, 1998) as a non-linear extension of a logistic regression estimator of the hazard rates (Efron, 1988) that arises naturally as the discrete time implementation of the proportional hazards model (Collett, 1994). This neural network model of survival has proved to be very stable in monthly studies over follow-up periods of several years, releasing the proportionality of the hazards assumption and fitting non-linear effects (Laurentiis *et al*, 1994, Biganzoli *et al*, 1998, Lisboa *et al*, 2000b). It also generates a prognostic index that can be interpreted in the same way as for the proportional hazards model, and is amenable to regularisation within the evidence framework of MacKay (MacKay, 1992) provided that account is taken of the highly skewed target distributions over time arising from very low hazard ratios in the time intervals (Lisboa *et al*, 2000b).

The main reported application of these models is mortality and recurrence following surgery for breast cancer. The practical importance of this work is to inform clinicians and patients regarding treatment. An early study of uncensored data for 5-year mortality from breast cancer and colorectal cancer, showed that the MLP predictions have a significantly better AUROC in each case than assigning each patient to the average survival for patients in the same TNM stage (Burke *et al*, 1997). A later study comparing the MLP with LogR also for mortality prediction,

gav comparable predictive performance for the two methods and indicated that good prediction was possible while omitting nodal status (Lundin *et al,* 1999). Two separate analyses of breast cancer recurrence used non-proportional hazards models where a conventional MLP replaces the linear risk score in Cox regression. One study found no significant difference in predictive accuracy of relapse within staggered fixed time periods (Ripley *et al,* 1998). The other study modelled contra-lateral recurrence of breast cancer, focusing on the structure of the prognostic index generated from the neural network and cross-checking it with stratified Cox regression (Mariani *et al,* 1997). This showed non-linear interactions between covariates which merit further clinical analysis. For example, low oestrogen was found to be protective in most patients, as is generally believed to be the case, but it appears to have the reverse effect in patients aged 45 or less. And, progesterone levels showed an interaction with histology, predicting a greater hazard of contra-lateral recurrence for patients with various types of ductal carcinoma, but a lower hazard for those with lobular carcinoma.

Overall, neural networks have not been clearly demonstrated improve upon the predictive power of proportional hazards, in breast as in other forms of cancer (Groves *et al,* 1999). However, an important conjecture is that classical and neural models of survival should be used as complementary, rather than rival tools (Biganzoli *et al,* 1998). In particular, one may feed low-order interaction terms onto the other, and both need to look more closely at common causes of error, namely the omission of key prognostic variables and the categorization of continuous scales (Schmoor and Schumacher, 1997). This is especially important in view of the remarks made in the methodology section concerning variable selection with neural network models.

2.2. Critical care monitoring

The few papers related to critical care listed in Table 5 are concerned with peri- and neo-nates. Predicting length-of-stay is increasingly important, and it is the focus of a comparative study of the MLP and multivariate linear regression (MLR) (Zernikow *et al,* 1999). A total of 40 first-day-of-life items were made available, together with the date of discharge. The paper concludes that even first-day-of-life data may contain sufficient information to usefully predict individual length of stay.

An earlier study to predict the likelihood of intra-cranial haemorrhage in pre-term neonates, with gestational age less than 32 weeks and birth weight below 1500g (Zernikow *et al,* 1998), again found the MLP to be superior to a benchmark linear model, this time LogR. This was on the basis of the AUROC, as well as the sensitivity measured over a range of clinically relevant specificities from 75% to 95 % in 5% steps. However, the neural network used 13 variables where stepwise LogR identified only 5. Given that stepwise regression is commonly found to select too many variables and hence be prone to over-fitting, and neither model was regularised, it is possible that the results reported are optimistic. An important related area is neonatal monitoring of fetal distress. Loss of oxygenation to the brain during labour results in acidic traces that can be monitored by the pH of the umbilical artery (Stock *et al,* 1994). This study lists the performances of networks with increasing complexity, without regularisation, showing the effect of over-fitting. The paper also raises the issue of skewed class label distributions, which is considered again in section 4.

Features extracted from EEGs have also been used to make statements about levels of abnormality, with potential for use as an early warning system in the paediatric intensive care unit. An example of this is an expert system with neural network and fuzzy logic elements, which was found to agree with an expert to within two from seven possible abnormality levels (Si *et al,* 1998). While all of these studies point towards potential for automated decision support systems

for an array of critical care applications, further evidence is still required from more widespread validation involving multi-centre trials, before these systems can be considered for routine use.

APACHE II is a severity of injury score in current use in intensive care units, although evaluations of its impact on patient outcome has not been uncontroversial. In particular, a multi-centre prospective cohort study involving 2,962 patients with one year follow-up showed that the severity of injury index does not always correlate well with measures of the extent of disability a year after discharge (Thornhill *et al*, 2000). This is an example of performance assessment not correlating well with patient benefit, instead showing greater than expected disability levels in patients admitted with an apparently mild head injury. Other lifestyle factors also had a significant influence on outcome. A direct comparison was carried out between predictions of mortality status at discharge for 8,796 patients from adult intensive care units, made by an unregularised MLP and a LogR classifier combining the APACHE II score, post-emergency operation status and a disease category coefficient (Wong and Young, 1999). The network inputs consisted of the 12 physiological variables that go into the calculation of the APACHE II score, together with post-emergency operation status, the patient's age and chronic health history The results show comparable accuracy for the two methods, on the basis of the Lemeshow-Hosmer χ^2 statistic typically used to assess mortality predictions by grouping data into predicted mortality groups in 10% intervals. This method highlights differences in calibration between the two methods, showing that APACHE II was closer to the calibration line in the mid-range and the MLP in the upper mortality range. This means that the meaning of a prediction of a particular percentage of deaths is over-estimated by APACHE when it is small and under-estimated when it is large, whereas the MLP over-estimates everywhere except in the top range.

2.3. Cardiolovascular medicine

Early diagnosis of acute myocardial infarction (AMI) in patients presenting at emergency wards suffering from severe chest pain, has attracted considerable interest for the application of pattern recognition. The main intention is to use biochemical markers from a blood sample, to predict the outcome of protein measurements whose results take several hours to obtain. The earlier AMI is detected so the sooner blood-thinning medication may be administered at a time when it has the largest effect in reducing the severity of the resulting damage to heart muscle. Several studies focus on this theme.

In an early study, Ellenius *et al*, 1997 followed-up the diagnosis of a patient with a minor AMI from the time of infarct, by monitoring the rise in the concentration of biochemical markers and identifying the stage at which the MLP, and each of three expert clinicians, could confirm the diagnosis. This unusual approach to system evaluation showed the model detecting AMI and later predicting the size of the infarct, simultaneously with the earliest firm indications by the experts.

A large-scale study of automated interpretation of 12-lead electrocardiograms for detection of AMI, was carried out with a cohort of patients presenting to a single hospital over a 5-year period, comprising 1,120 confirmed cases and 10,452 controls. A 20 s trace was represented by six automatically generated ST-T measurements from each of the 12 leads, providing inputs to 72 input units of a MLP with a single hidden layer (Hedén *et al*, 1997), controlled for over-training by early stopping tuned with eight-fold cross-validation. The same cross-validation procedure was used for performance estimation, showing a 15.5% (CI 12.4-18.6%) sensitivity improvement over rule-based criteria used by computersized electrocardiographs at the emergency department, with 95% specificity. A smaller improvement of 10.5% (CI 7.2-13.6%) was found over the detection rate for AMI by expert cardiologists, who were restricted to reading the ECGs in the

absence of contextual information such as personal data and clinical findings at the initial examination, reducing the specificity to 86.3%. The comparison with the computerised rule-based criteria represent a 50% increase in boost-factor from approximately six to nine, whereas in the comparison with clinicians the change is by 20% at a likelihood gain of around 3.5. The improvements deteriorate at higher specificities, but these results have nevertheless been reported enthusiastically in medical journals (Josefson, 1997, Fricker, 1997). The authors point out that the system would be of particular value to junior doctors in emergency rooms, and the next stage is to integrate additional contextual information such as patient history and findings of clinical investigations. Some of these ideas are pursued further in the work described next.

An advanced methodological study of AMI detection in emergency departments with neural networks, comprises a sequence of papers by Baxt and collaborators. Early papers to optimise the accuracy of the neural network predictions (Baxt, 1992) were followed by a careful analysis of the effects of individual clinical inputs on the network decision (Baxt, 1994), and the application of rigorous practical methodologies for sensitivity analysis (Baxt and White, 1995). Of particular interest is the use of the bootstrap to correct for finite-size effects, causing bias in the sensitivity estimates derived from the training data, a sample with 706 observations. This bias is significant enough to change the rank-order of importance of the clinical inputs.

The analysis of input effects by calculating bias-corrected sensitivities in Baxt and White, 1995, ranked new variables higher than certain indicators commonly used by expert clinicians. The resulting model is consistent with another study of variable selection for the prediction of AMI (Dreiseitl *et al*, 1999) comparing LogR, Bayesian neural networks (Neal, 1996) and rough sets. Several variable selection methods suited to each modelling approach were also applied to a set of 500 records, selecting from 43 variables. Multiple variable selection runs were carried out with a training data consisting of 335 patient records, optimising the results for a test set comprising the remaining 165 records. Only one variable, ST elevation, was selected by all methods. This is recognised as clinically very relevant, but three other variables commonly used by experts were not selected by any of the models, namely history of diabetes mellitus, severe chest pain and pain duration. The results of both studies are consistent with respect to these variables, suggesting that while they may be sensitive to AMI, they have poor specificity.

The initial studies were followed by a prospective comparison between the detection rates by cardiologists and the MLP for a cohort of 1,070 patients aged 18 and over presenting with anterior chest pain, again, to a single hospital (Baxt and Skora, 1996). This is an observational study in which the clinicians and decision system separately reviewed the same patients, therefore it was not listed in PUBMED as a Controlled Trial. It does not qualify as an evaluation of clinician performance, since the trial does not involve clinicians in the loop, or of patient outcome, as the decision support was not used to inform clinical decisions. Nevertheless, it claims arguably the largest increase in performance by a neural network system compared with expert clinicians, boosting the likelihood of making a correct decision by factors of 24 and 4.7, respectively, over the prevalence.

An earlier, multi-centre trial involving emergency departments in six hospitals, compared three quite different modelling structures for classification, namely rule induction, LogR and the MLP, for the prediction of acute cardiac ischaemia (ACI), comprising AMI and unstable angina pectoris, from 8 variables available within the first 10 minutes of emergency care (Selker *et al*, 1995). The variables represent patient history, together with features extracted from a clinical examination and an electrocardiogram. The MLP had noticeably poorer calibration than the other methods. It also suffers from being more difficult to interpret. It was concluded from this study that the choice of database decision support method should be made on the basis of specific

application needs rather than on the premise that any of the methods tried is intrinsically more powerful than the others. Nevertheless, the results obtained indicate that an important limitation of the predictive performance for this clinical outcome is the availability of reliable data, rather than the need for more algorithmic development.

An altogether different application is to predict the likelihood of patients developing transiet myocardial ischaemia during a period with ambulatory Holter monitoring, using parameters form a previously recorded 12-lead resting ECG (Polak *et al,* 1997). This is an example of a study where the MLP trained by back-propagation was out-performed by a linear discriminants analysis (LDA) and an alternative model to the MLP, the adaptive logic network. The poor performance by the MLP, which in principle has enough flexibility to emulate the other two, could be due to substantial over-fitting. The adaptive logic network is an early derivative of the MLP where all nodes are forced towards the hard-threshold limit, thus assuming a Boolean logical function. Constraints can then be applied to explicitly control qualitative features of the model, for instance the convexity of the decision boundaries.

Turning now to diagnostic radiology, a study of myocardial images to quantify coronary heart defects from perfusion scintigrams consisted of developing a classifier with images from 135 patients from one hospital, and testing it in another (Lindahl *et al,* 2000). This external validation is a key stage in the evaluation of improvements in clinician performance arising from the deployment of a decision support system, since anything from the quality of equipment to acquisition protocols may vary between clinical centres ostensibly performing equivalent tasks. The sensitivity of the MLP was found to exceed that of alternative rule-based detection algorithms including some derived from the Cedars-Emory quantitative analysis software, CEqual (de Sutter *et al,* 2000). In an earlier study of the influence of decision support on the interpretation of bull's-eye scintigrams by clinicians, the images were independently classified by three experts twice with, and twice without advice from the neural network (Lindhal *et al,* 1999). Overall, there was significantly less inter- and intra-observer variability in detection of presence against absence of coronary disease and classification into two from four categories than without support, with corresponding increases in diagnostic accuracy measured by the AUROC.

Georgiadis *et al,* 1995, is another rare study aiming at a systematic characterisation of intra-observer, inter-observer and intra-subject variability. In this carefully conducted feasibility study of automated embolus detection in patients with prosthetic heart valves, also derived from Doppler ultra-sound measurements, no significant difference was found between the microembolic counts of different observers, among three separate counts by the same observer, or between human observers and a MLP. In addition, repeat examinations of the same patient were also consistent, indicating that the detection of microembolic signals in this patient cohort is a reproducible technique.

In a separate application, Doppler ultra-sound waveforms were used to detect proximal, distal and multi-segmental stenosis at the site of the common femoral artery (Smith *et al,* 1996). Promising results were obtained with an MLP by sub-sampling the waveforms of blood-velocity over time, but only in the discrimination of healthy from diseased and not into the four initial categories. This approach was significantly more accurate than a Bayesian classifier using PCA scores. In an unrelated study, Goodenday *et al,* 2000 applied an image recognition network directly onto perfusion scintigrams. The network used the principles of overlapping localised receptive fields familiar from the Neocognitron, but even then difficulties were experienced with changes in position and scaling between the images.

2.4.Other applications

The remaining publications describing medical neural network applications listed in PUBMED Randomized Controlled Trials or Clinical Trials are in the areas listed below:

- Diagnosis (Pesonen, 1997, Goodey *et al,* 2000, Sonke *et al,* 2000, Grus and Augustin, 1999, Kemeny *et al,* 1999, Smith *et al,* 1998, Leon and Lorini, 1997, Gurgen *et al,* 1995, Kimberley *et al,* 1994)

- Outcome of treatment (Modai *et al,* 1996, Dombi *et al,* 1995, Michaels *et al,* 1998)

- Physiological measurement:

    - MEG (Gaetz,1998), visual evoked potentials (Liestritz *et al,* 1999) and EEG (Anderson *et al,* 1998, Baumgart-Schmitt *et al,* 1998 *et al,* 1997, Winterer *et al,* 1998, Heinrich *et al,* 1999, Grozinger *et al,* 1998, Guterman *et al,* 1996)
    - Other measuremens (Taktak *et al,* 2000, Liang *et al,* 2000, Chen *et al,* 2000, Tafeit *et al,* 1999, Kol *et al,* 1995, Barnhill *et al,* 1995)

- Radiology (Bakken *et al,* 1999, Park *et al,* 1998, Szabo *et al,* 1996, Horwitz *et al,* 1995)

- Pharmacokinetics (Chen *et al,* 1999)

- Physical medicine and rehabilitation (Chang *et al,* 2000, Riess and Abba, 2000, Wu and Su, 2000, Savelberg and De Lange, 1999, Simpson and Levine, 1999, Patterson and Draper, 1998, Kiani *et al,* 1997, Abbas and Triolo, 1997, Chang *et al,* 1997)

A useful diagnostic benchmark for decision support systems is acute abdominal pain. Here, neither neural networks (Bounds *et al,* 1988, Horace Mann and Brown, 1991, Pesonen, 1997) nor machine learning methods (Ohmann *et al,* 1996) have had much impact. Unregularised MLPs were benchmarked against linear discriminants, LogR and cluster analysis, with the outcome that none of the methods applied were found to be superior to each other, or to independent Bayes. Study design was typical of early prototype applications and much more divorced from integrating into the clinical decision process than de Dombal's approach. In particular, one of the most revealing aspects of de Dombal's system is seldom built into today's prototypes. This consists of two additional steps in the feedback to the clinician. The first is a comparison between the model outcomes, and two ranked diagnoses indicated by the clinician. De Dombal's advisory system uses this information to advise further tests that are the most likely to resolve the discrepancy between the two sets of clinical decisions, providing the required evidence to re-assess the clinical decision. Secondly, the model suggests a list of possible rare diseases, to alert the clinician in obscure cases, obviating the need to carry 'encyclopaedic' knowledge about the wide range of possible manifestations of a disease. It is somewhat surprising that this diagnostic system, which in a large multi-centre study reduced the residual diagnostic error for this notoriously difficult condition by 40% nearly two decades ago (de Dombal *et al,* 1993) and integrated well with routine clinical practice, is not more widely known.

Response to treatment was the subject of a RCT comparing interventions by expert clinicians with those guided by a decision support system, using a different architecture from those met previously in this review, the Adaptive Resonance Theory (ART) network. This is potentially the model of choice for generic medical applications, since a common expectation of an 'intelligent

system' would include the need for incremental learning and transparency of interpretation, at least by providing access to category prototypes. Treatment recommendations were made for 26 schizophrenic and 28 unipolar depressed patients, by randomly allocating to choice of intervention to the consensus of two senior psychiatrists, or the recommendations made by an ART network trained on a set of 211 historical cases all of whom had improved during 8 weeks of treatment (Modai *et al,* 1996). The recommendations from ART were similar to those from practicing psychiatrists, with no contra-indicated treatments suggested, although two incomplete treatment suggestions were removed from the study. Similar hospital lengths of stay were experienced by both sets of patients.

The length of stay of rib fracture patients and their ICU days and mortality were also predicted, training on 522 patients and testing on a further 58 patients, both sets randomly allocated from a pool of 580 cases. Although this is a retrospective observational study, it was listed under RCT. Several MLPs were trained which showed encouraging predictive accuracy (Dombi *et al,* 1995). More importantly, their weights were examined to identify main input effects, raising hypotheses which could be followed-up by conventional statistical modelling.

A particularly interesting controlled trial is a comparative assessment of three primary-to-secondary care referral strategies for patients with third molars (Goodey *et al,* 2000). This is a true multi-centre control study, involving 32 primary care dental practitioners who were randomly allocated to current practice (the control group), neural network-based decision support, or a paper-based clinical decision algorithm. The referrals of 107 patients were assessed by a panel of experts against a gold standard consisting of criteria from the National Institutes of Health. The control group's figures for accuracy, sensitivity and specificity (0.83, 0.97, 0.22) display significantly better accuracy and sensitivity than either the neural network (0.67, 0.56, 0.79) or the clinical algorithm (0.73, 0.56, 0.93), albeit with much poorer specificity. The study recommended integration into primary care of paper-based guidance, as the best overall compromise. Separating the effect of prevalence from the raw ROC figures, the boosting factors for the control, neural network and clinical table groups were 1.2, 2.7 and 8.0, respectively, indicating a greater discriminatory effect as a result of the selected referral guidance approach.

The remaining clinical trials listed in Table 2, describe applications to medical domains outside the focus of this review. The neural network methodologies employed in electro-physiological measurement mostly rely on the standard MLP applied to signal representations generated by principal components analysis or the coefficients of auto-regressive models, with the exception of one study using wavelet networks. (Heinrich *et al,* 1999). The choice of medical application reflects the increasing use of EEG and evoked potentials for diagnosis and monitoring of a wide range of conditions. While the predictive accuracies reported are encouraging, few systems have been tried in routine use in a clinical setting, notably one reported elsewhere for sleep analysis (Davies *et al,* 1999).

Other physiological measurements also generate complex signals whose interpretation is fertile ground for decision support. Reported CTs address areas of current practical relevance, including monitoring of fetal distress (Kol *et al,* 1995) and artefact detection in SaO_2 and $TcPO_2$ (Taktak *et al,* 2000), analysis electrogastrograms (EGG) to detect delayed gastric emptying (Liang *et al,* 2000, Chen *et al,* 2000), visualisation of sub-cutaneous fat in patients with diabetes mellitus (Tafeit *et al,* 1999), and establishing a correlation between bone demineralisation measured by x-ray absorptiometry and a composite serum index (Barnhill *et al,* 1995).

The emphasis in the clinical trials with neural networks listed under radiology is on functional, rather than morphological, imaging. Relatively new modalities such as PET (Szabo *et al,* 1996,

Horwitz *et al,* 1995) and MRS (Bakken *et al,* 1999, Park *et al,* 1998) have considerable potential to characterise metabolic characteristics of *in vivo* tissue, but often result in complex signals with difficult quantitation. In particular, the analysis of *in vivo* MRS is currently undergoing rapid development, including source identification with non-linear signal processing (Lee et al, 2000).

In pharmacokinetics, a RCT addressed the wide inter- and intra-subject variability that hamper predictions of drug concentrations in blood. A retrospective study of tracolimus levels in 32 liver transplant patients (Chen *et al,* 1999) concluded that the blood concentration of this anti-rejection drug is accurately predicted by a MLP. This paper adopts a genetic algorithm as a generic methodology to optimise model design through variable selection.

A number of CTs in the area of physical medicine and rehabilitation cluster around the assessment of high-dimensional complex signals used, for instance, in gait analysis (Chang *et al,* 2000, Wu and Su, 2000), non-linear control of functional electrical stimulation (Riess and Abbas, 2000, Abbas and Triolo, 1997 and Chang *et al,* 1997). The remaining studies relate to non-linear mapping of insole pressure patterns into a grouped-reaction force (Savelberg *et al,* 1999), activity detection in ambulatory monitoring (Kiani *et al,* 1997), and wheelchair propulsion (Patterson and Draper, 1998) and navigation (Simpson and Levine, 1999).

3. Impact of medical decision support with artificial neural networks

There are seven trials of neural network-based decision support systems (CDSS), involving one to twenty one thousand patients. The most impact to date has been in cervical cytology, where PAPNET has arguably established a new standard for sensitivity in detection of dysplasia (Prismatic team, 1999, Doornewaard *et al,* 1999, Mango and Valente, 1998). The other large studies show promise to identify patients at low risk of lymph node spreading in prostate cancer (Gamito *et al,* 2000), predicting lengh-to-stay of neonates in pediatric intensive care units (Zernikow *et al,* 1998), early detection of acute myocardial infarction (Selker *et al,* 1995) and prediction of transient ischaemia during ambulatory ECG monitoring (Polak *et al,* 1997).

The studies of neural network assisted cytology compare with conventional cytological screening as the control intervention (Prismatic team, 1999, Doornewaard *et al,* 1999, Mango and Valente, 1998). Other true control trials are for response to treatment of head & neck carcinoma (Bryce *et al,* 1998), treatment advice in schizophrenia and depression (Modai *et al,* 1996) and referral from primary to secondary care in dental practice (Goodey *et al,* 2000). All of these studies advise the use of neural networks in a clinical supporting rôle, on the basis of improvements in patient performance, rather than by direct evaluation of changes in patient outcome.

Almost all of the of the comparative trials are prospective and assess the performance of users external to the designers of the CDSS, but only six of them are multicentre trials (Goodey *et al,* 2000, Gamito *et al,* 2000, Doornewaard *et al,* 1999, Mango and Valente, 1998, Prismatic team, 1999, Selker *et al,* 1995). Frequently, the criteria for RCT listing appears to have been random allocation to design and validation datasets, rather than the mandatory allocation to active or control interventions. For neural networks to be taken seriously by clinicians as inference models, it is essential to integrate them into systems that stand-up to the gold standard of clinical evaluation, namely multi-centre RCTs. This is widely regarded to be a key milestone for the evaluation of any type of medical decision support, since it validates against the vagaries of inter-patient and inter-centre variability.

Yet, almost all of the publications listed in the survey are observational studies with retrospective data, pitting the multi-layer perceptron (MLP) against multivariate logistic regression (LogR). In

most of the studies listed in PUBMED as clinical trials, there is no attempt to separate a design dataset, used for training and parameter tuning, or testing, from a validation set used for performance estimation. This is know to result in optimistic assessments of performance (Tibshirani, 1996, Efron and Tibshirani, 1997) and undermines claims to generality, that is to say generalisation to out-of-sample data.

As with the evaluation of any new intervention, decision systems must eventually be assessed by a trial that is patient, rather than clinician based. However, the need for further external evaluation of patient benefit must not detract from the substantial achievements made with neural networks in the first two decades since they became widely known, which is relatively recently in comparison with symbolic systems and even more with so compared with linear statistical inference. It is also notable that evidence is continuously mounting towards the power of quantitative modelling to evidence difficult decisions in a very wide range of clinical and medical laboratory applications. Moreover, neural networks are showing their worth in exploratory studies to uncover important un-modelled non-linear interactions, and to provide effective visualisation of complex high-dimensional data. For instance, self-organised maps (SOM) may gradually fill a niche in the specific characterisation of subtle transitions in disease indicators marking qualitative changes in the grading of disease, for example in tumours.

With regard to methodological issues, in many of the applications reported there are further steps to clarify the model structure and improve the robustness of the conclusions that add little additional complexity to the study. In particular, regularisation of the objective function, for instance using weight decay (Bishop, 1995, Ripley, 1996), is a straightforward way to protect against the major curse of universal non-linear maps, which is over-fitting of the data (Astion *et al*, 1993). Another key element of all forms of statistical analysis, whether linear or not, is a close inspection of the sensitivity of the model response to variations in the value of the predictor variables. This has to be done with care, since it will influence model selection (Tibshirani and Knight, 1999). An additional element of complexity involving systematic re-sampling (Baxt and White, 1995) serves to remove bias from small sample estimates of the response sensitivity of the model, and the same bootstrap process will also substantially improve estimates of generality (Tibshirani, 1996, Efron and Tibshirani, 1997). In performance evaluation for differential diagnosis, the Receiver Operating Characteristic (Hanley and McNeil, 1982) is the *de facto* standard, but it is regrettable that scant attention is given to the usually skewed nature of the data, either by reporting the boost afforded over prevalence, or during training, by directly maximising predictive power rather than accuracy (Lisboa *et al*, 2000b).

Neural network applications in oncology have been formally reviewed in a statistical journal, with critical results (Schwartzer *et al*, 2000). In this paper, uncritical use of the very flexibility that underpins the non-linear mapping capabilities of neural networks, is shown to generate implausible functions resulting in under-estimation of the misclassification probabilities leading to 'exaggerated claims' the potential for neural network models for diagnosis and prognosis. The paper identifies frequently made mistakes in applications of artificial neural networks, including over-optimistic claims of generalisation performance, training large networks with small data samples, use of inadequate statistical benchmarks and lack of significance in the comparison of performance results for neural networks against alternative, often simpler, statistical or rule-based models. It is also pointed out that survival models have been too naïve, sometimes consisting of single time-point models which ignore censorship.

Other reasons for the poor take-up of decision support in routine practice are listed in a more general commentary about statistical prognostic models (Wyatt and Altman, 1995). A common criticism is that clinical factors modelled may have little impact on decisions about treatment, use

model structures that lack credibility because they violate well-established clinical precepts of cause-and-effect pathways, or are insufficiently validated. The immediate conclusion is that many published models report prototypes but need further research before clinical adoption (Haynes, 1990). This remains current for medically related neural network systems, where good practice could be improved by attending to certain methodological considerations that are summarised in the next section.

4. Methodological considerations for neural networks in medical applications

Artificial neural networks are characterised by their ability to model complex systems but several shortcomings of their use arise from lack of robustness in controlling this flexibility. In a parallel development, the analysis of increasing complexity in statistical inference led to the proposal of baseline criteria for good practice in the design of clinically relevant models (Concato *et al*, 1993, Altman and Royston, 2000). These recommendations are now incorporated into a blueprint for the design of complex decision systems, which is reviewed here with reference to neural networks in medicine, drawing from the critical analysis in the preceding sections.

4.1 Clarify the purpose of the study

Even a cursory overview of medical neural network prototypes indicates how they are taken to be synonymous with inference. This is not necessary, nor is it where most of the benefit of these methods can be derived, when they are set alongside the plethora of available statistical and knowledge-based methodologies. Altman and Royston, 2000 recognise different categories of prognostic studies, and their comments apply also to diagnostic modelling:

- Pragmatic studies have the purpose of providing direct advice that will affect decisions about treatment. It is, therefore, essential to pre-specify the performance levels at which this support becomes useful, as well as to quantify discriminant power in the most robust possible way. These studies should ideally fulfil the requirements for Phase II exploratory trials, by involving multiple clinical centres to ensure that the performance estimates obtained may be used to specify, not just the protocol but also the sample sizes required for a Phase III RCT. Sample size estimation for neural networks is an area of current research, but practical methods have been published (Chan *et al*, 1999).
- Exploratory studies aim to generate new understanding about the condition and may consist of visualisation of the generation of hypothesis for inference models. One example is to generate hypothesis about important interactions between covariates, which can be tested with standard medical statistical methodologies. Another is to understand where the variables selected by the model fit in relation to prior knowledge from the medical domain.

Above all, it is essential to be clear about the purpose of the study, and specify in advance what aspects are expected to the valuable to support subsequent studies.

4.2. Model design

In the application of neural networks, good practice in model design is critical. There are standard tools to control for over-fitting, some of which also help in variable selection.

4.2.1. Network regularisation

The vast majority of applications reviewed here use what could be termed first generation neural networks. These rely for their generalisation on a parsimonious design with few hidden nodes,

and early stopping by reference to a test dataset as the means to prevent overtraining (Caruana et al, 2001). However, using too few hidden nodes can reduce the variance of the predictions at the cost of significant bias due to insufficient complexity in the model. It follows that simpler interpolation is achieved by using many hidden nodes but imposing direct regularisation of the objective function, the simplest of which would be weight decay (Bishop, 1995, Ripley, 1996).

A development of this approach is to use Bayesian estimation of the hyperparameters, usually by assuming broad unimodal normal distributions that are well represented by just their mode, called the evidence framework (Mackay, 1992, Bishop, 1995). This is a practical and robust methodology for neural network design, which has the advantage of providing control for large networks with reproducible results that do not require extensive hand-tuning (Husmeier *et al*, 1999). However, care must be taken to ensure calibration for unbalanced data, which is typical of medical applications and especially acute in survival modelling (Lisboa, 2001). The third generation of neural networks consists of implementing the full Bayesian integrations by importance sampling with efficient Markov Chain Monte Carlo algorithms (Husmeier et al, 1999, Neal, 1996). This has the further advantage of more accurately predicting distributions of the posterior, or of regression inferences, even in the presence of skewed data.

Further developments of neural networks for inference optimise radically different principles from penalised log-likelihood. For instance, support vector machines are based on the precepts of structural risk minimisation, rather then directly estimating the empirical error. They are reported to provide greater robustness for small sample sizes, features that are especially relevant for medical applications (Cherkassky and Muller, 1998). Another research direction for medical neural networks aims to identify when inferences should be made at all, as they are not reliable for new observations that lie outside of the multi-dimensional support of the design data. This is the area of novelty detection (Campbell and Bennett, 2001), where methodologies are being developed with a rôle that is orthogonal to the response modulation by the confidence terms in the evidence approximation.

In a parallel development to the use of neural networks for inference, models for visualisation and clustering have also been embedded into increasingly rigorous theoretical statistical frameworks. Visualisation is a powerful tool to gain understanding of the phenomenology of disease, although its use tends to be confined to decision support for diagnosis or prognosis. An example of a flexible neural network method for visualising the distribution of indicator variables over the decision space is the Growing Cell Structure technique (Walker *et al*, 1999). This is a self-organised model whose usefulness in visualisation is matched by good performance in clustering an externally imposed category label, mirroring similar characteristics for the conventional SOM (Dreiseitl *et al*, 1999). An implementation of a topographic clustering algorithm regularised within a Bayesian framework similar to the evidence approximation, is the Generative Topographic Mapping model (Bishop *et al*, 1998). Other generative approaches have been developed also for signal processing, specifically for blind signal separation (Cardoso, 1999) which rapidly becoming a *de facto* standard in advanced signal processing and is extensively applied to the detection of sources in electrophysiological measurements (Lisboa *et al*, 2000a), as well as the identification of sources in static complex signals (Lee *et al*, 2000).

4.2.2. Variable selection

It is also important to consider that the benefit of the capability for universal function approximation is realised only when there are unsuspected non-linear interactions between the predictor variables, for instance where the precise form is unknown or their effect changes in

complex ways in different areas of pattern space. This requires at once careful variable selection and detailed interpretation of the relationships between the covariates and the model predictions.

The GUSTO trial also applied a Bayesian regularisation framework using the evidence approximation (MacKay, 1992) with separate weight decay regulators for each predictor variable (MacKay, 1995, Neal, 1996). This is known as Automatic Relevance Determination (ARD) because large values of weight decay hyperparameter indicate a narrow distribution of the posterior probabilities for the weights around the prior value of zero, while smaller values relax the posterior distribution of the weights further away from the prior, indicating a covariate that is effective in discrimination. This provides useful indicators to use in variable selection which, for neural networks, is notoriously unstable resulting in multiple candidate models that also differ from those selected by linear statistics (Dreiseitl *et al*, 1999).

A useful yardstick for the estimation of sample sizes for exploratory studies is to require the number of observations to be in the range of 5 to 10 times the number of available (rather than selected) covariates. This is because given enough random variables, there will be some that correlate with any given sequence of labels. In the case of artificial neural networks, the effective number of degrees of freedom is not straightforward to estimate, but this can be done with the evidence approximation using factors related to the regularisation hyper-parameters (MacKay, 1992, 1995). Recently other, more generic methods are also being proposed for estimation of the effective number of parameters in linear and non-linear models (Tibshirani and Knight, 1999).

4.3. Validation: support for learned intermediaries

One of the most important legal doctrines that apply to the use of decision support systems in medicine is the doctrine of learned intermediaries (Braham and Wyatt, 1989). This requires clinicians to understand the operation of the system well enough to be able to take responsibility for the results of its use. From a system designer's perspective, the need to explain the model is a necessary step in validation, by scrutinising the agreement between the inference model prior clinical understanding of the data (Lisboa, 2001). Similarly, in exploratory studies the most useful result is often, not whether a particular methodology increases diagnostic or prognostic accuracy, but why it appears to do so. In general, whatever the model accuracy in statistical terms, doctors will be reluctant to use it to inform their patient management unless they believe in the model and its predictions (Altman and Royston, 2000). Therefore, it is essential to test all relevant clinical data for inclusion in the model, ensuring that data are accessible and reliable, and preferably already routinely acquired. Other considerations are also important in maintaining the integrity of the data representation, for instance avoiding the introduction of arbitrary thresholds for continuous variables, or ignoring censorship in prognostic models, either of which is prone to introduce bias in the model predictions.

A useful tool to constructively compare linear and non-linear models and to explain the influence of covariates on model response, is to carefully analyse the profiles of input variable distributions in prognostic groups, or calibration segments in diagnosis. Another useful step is to analyse the input-response characteristics of the selected variables using sensitivity analysis. A caveat of this approach is that small sample effects introduce considerable bias into naïve estimates of input effects, but established methods to compensate for that bias already exist (Baxt and White, 1995). In this paper, the bias-corrected sensitivities of a MLP for AMI prediction show new significant variables, in rales and jugular venous distention. All significant effects are in directions which accord with clinical expectation. Several of the predictor variables were found to have bi-modal distributions of the re-sampled bootstrap statistics of input effects, indicating the presence of contextual effects characterised by the other variables. The discovery of unexpected interactions

between covariates is key to understanding the predictive value of neural networks compared with alternative methodologies. Some of these interactions may be low-order and can, therefore, be ported into linear-in-the-parameters statistical models that already carry the confidence of clinicians, thus gradually enhancing existing tools in a controlled manner. An alternative approach where theoretical foundations are the subject of active research is rule extraction from trained neural networks (Hayashi *et al*, 2000).

4.4. Benchmarking against a suitable alternative

Another common regulatory requirement is to demonstrate that a new technology performs at least as well as an alternative method which is substantially equivalent to that used in systems that have been previously certified. In the case of the MLP, the obvious benchmarks are MLR in regression, and LogR for classification, and for survival modelling it is the proportional hazards model. In order to gain the confidence of the statistical community it may be necessary to demonstrate parity with linear-in-the-parameters models, where non-linearities are explicitly coded-in. This leaves out only unexpected interactions between covariates, whose presence in small samples of typically noisy medical data, may be difficult to demonstrate.

One of the most frequent criticisms of neural network prototypes is the excessive claims of performance benefit made without apparent regard for the significance of small differences between the performance achieved with alternative methods (Schmoor and Schumacher, 1997). For inference models, this is not a justifiable omission since accepted tests exist to monitor the significance of differences occurring by chance, such as McNemar's test on discordant pairs (Schmoor and Schumacher, 1997 and Ripley, 1996).

4.5. Robustness in performance evaluation

The sources of uncertainty that reduce the ability to maintain performance results from one patient cohort to another are many and varied, including within-patient variation, between-patient variation, case mix differences which involve different confounding factors for each cohort, as well as instrumentation and protocol differences between clinical centres. Consequently, studies need to clearly distinguish between:

- internal validation, where a design sample is used to train and tune the model parameters (test sample) but new data (validation sample) are used for performance estimation,
- temporal validation using later data from the same clinical centre, and
- external validation where the data come from entirely different clinical centres not involved in the model design.

Whereas modelling studies start use retrospective data, phase II exploratory studies normally require prospective data and phase III clinical trials require the prospective application of the system to large numbers of subjects across multiple centres.

When sample sizes are small, cross-validation is sometimes used for performance estimation, although a more robust method would be to use the bootstrap (Tibshirani, 1996, Efron and Tibshirani, 1997, Jain *et al*, 2000). It can be emphasised that optimistic performance claims are a big barrier to the take-up of neural networks by the medical community. Careful variable selection, together with appropriate regularisation, benchmarking and robust performance estimation will make research findings less impressive, but more meaningful.

One aspect of performance evaluation that is routinely observed in the papers reviewed, is the analysis of results within the Receiver Operating Characteristic (ROC) framework (Hanley and McNeil, 1982). What is less commonly done is to separate out also the effect of prevalence of different diseases (Hilden, 2000), which significantly affects estimates of accuracy. This is important because poor calibration may be masked by a good AUROC, since the latter aggregates inferences made either side of the decision threshold (Lisboa *et al*, 2000b). Predictive power may be factorised by combining the sensitivity and specificity at the operating decision threshold to calculate the boost over the guessing line, shown in equation (1).

4.6. Comparative trials

The design of comparative trials to evaluate the changes to expert intervention with and without access to a decision support system, is fraught with difficulties (Ohmann *et al*, 1999, Hunt *et al*, 1998). For example, in a controlled prognostic trial of the management of acute abdominal pain, involving 558 patients, the clinician's own diagnostic performance was significantly enhanced (de Dombal *et al*, 1974). In particular, the proportion of appendices that perforated before operation fell from 36% to 4%. A later study with a further 295 patients concluded that after a two-month learning period the system proved more accurate in its diagnoses than unaided clinicians, and during the first five months of using the system the unaided clinician's performance rose from 73% to 84%, the so-called checklist effect, also known as the Hawthorne effect (Adams *et al*, 1986, Randolph 1999). This effect is important because it can obscure the contribution of decision support systems even when clinician performance is improved, in effect reducing the statistical power of the RCT. Adams *et al*, 1986 give guidelines for minimizing the checklist effect and to ascertain the significance of the results obtained, allowing for different patient statistics in the experimental and control groups as well as the pooling of multiple clinical centres. For a more complete discussion of the technical issues surrounding the design of RCTs, see Campbell *et al*, 2000.

Altman and Royston, 2000, also make a distinction between valuable studies and those that are merely valid. This harks back to the very specification of the purpose of the study and the relationship between inference making and its effect on patient management (Murray *et al*, 1993). In relation to the publications reviewed here, it is noteworthy that none dealt directly with changes in patient outcome, which do not always readily follow improvements of clinician performance (Johnston *et al*, 1994). These are factors which merit detailed consideration at the start of any study into clinical decision support, alongside the realisation that improvements may also be achieved, sometimes more effectively, by non-technological means. This is important because clinicians will typically accept the value of additional training more readily than equivalent benefits arising from technological aids. For instance, a redesign of the processes for the interpretation of radiographs in the emergency department was found to reduce tenfold the rate of false negative errors, such as missed fractures or foreign bodies, even from a starting base as 3% (Espinosa and Nolan, 2000). Taking an example where neural network-based decision support has been successful, the 42% reduction in UK mortality rates from cervical cancer between 1987 and 1997, has been achieved by across the board systematic improvements to care, covering call and recall systems, education and training for smear takers and cytologists, even an awareness of the need to audit the screening histories of all cases that pass through the detection filter to result in invasive cancer (Cuzick and Sasieni, 1999). For these reasons, it is necessary to appreciate the complete clinical picture before targeting the introduction of new computer-based systems.

5. Conclusions

Artificial neural networks have been identified alongside the bootstrap, Bayesian modelling using Gibbs sampling (Lunn *et al*, 2000), generalised additive models and CART as new trends in medical statistics (Altman, 2000). Although none of these methods is in widespread use, neural networks have had a clinical impact in specific areas, notably cervical cytology and early detection of AMI, where large-scale prospective multicentre studies have been carried out. More generally, rapid developments in instrumentation, communication and data storage, will ensure that increasingly complex signals will become routinely available in digital form. Coupled with pressures for greater accountability of clinicians and a shift towards systemic approaches to the management of medical error, these developments will ensure that the practical rôle for automated decision support in medicine will continue to grow (Wilson *et al*, 1995, Weingart *et al*, 2000, Barach and Small, 2000, Reason, 2000).

Neural network inference has been most useful for closely circumscribed tasks where there are significant interactions between covariates. This has applications in tasks that require attention focusing, such as the detection of a few abnormal cells in a large number of cells present in a slide. The range of prototypes already reported in the medical literature is evidence of the potential of intelligent medical instruments for multivariate prognostic or diagnostic inference, and to provide practical visualisation of high-dimensional signals. Besides their rôle in supporting evidence-based predictions and to reduce information overload, generic non-linear models are also useful in exploratory data analysis, by generating hypotheses about complex terms that may be integrated in a controlled manner into standard statistical models.

However, the potential for further use of statistics and pattern recognition in medicine, is not specific to neural networks. It is increasingly important to demonstrate good practice in their design and in verification and validation, and to benchmark them against structurally simpler models that have been appropriately optimised. Currently, the claims made in too many prototype studies are not robust. This is not unique to neural networks, indeed it is reminiscent of the early use of linear discriminants analysis (LDA) (Lachebruch, 1977) and, arguably, even of the general state of mainstream applications of statistical techniques in medicine (Altman and Goodman, 1994, Wyatt, 1995).

There are other important factors that limit the take-up of intelligent decision systems generally, namely the need to design systems that address real clinical needs, and which are more readily integrated into the routine data-management environment of the user (Potthof *et al*, 1988, Shortliffe, 1993). Achieving this has been the hallmark of the few successful neural network applications that have made it into routine clinical use.

Schwartz, 1970, predicted that by the year 2000 computers would 'have an entirely new rôle in medicine, acting as a powerful extension of the physician's intellect'. While this prediction is far from being realised for artificial intelligence tools in general clinical consultation, it has been true of developments in medical instrumentation. Over the last decade, inference-based decision support systems have started to emerge in routine clinical use on a significant scale. Some of these are purely statistical (APACHE II, Rowan *et al*, 1994, AAP, De Dombal, 1984, GLADYS, Davies and Owen, 1990), and PAPNET (Koss, 2000) involved first generation neural networks.

In the future, the rôle of computers in medicine will be substantially extended along several directions. These range from patient management, with the development of application service providers and knowledge-based decision support systems, through increased sophistication in electronic systems for data acquisition, storage and transmission, spurred-on by a gradual

acceptance of industry standards such as extensible mark-up languages, onto the emergence of radically new applications in telemedicine and self-care. In the global context of healthcare as a commodity, decision support is likely to become a necessity rather than an optional extra, just as advanced electronic instrumentation is today. Neural networks have a niche to carve in clinical decision support, but their success depends crucially on better integration with clinical protocols, together with an awareness of the need to combine different paradigms in order to produce the simplest and most transparent overall reasoning structure, and the will to evaluate this in a real clinical environment.

5. References

Abbas, J.J. and Triolo, R.J. Experimental evaluation of an adaptive feedforward controller for use in functional neuromuscular stimulation systems. *IEEE Rehabil Eng* 1997, 5(1):12-22.

Adams, I.D., Chan, M. Clifford, P.C., Cooke, W.M., Dallos, V., de Dombal, F.T., Edwards, M.H., Hancock, D.M., Hewett, D.J., McIntyre, N., Somerville, P.G., Spiegelhalter, D.J., Wellwood, J. and Wilson, D.H. Computer-aided diagnosis of acute abdominal pain: a multicentre study, *BMJ* 1986, 293:800-804.

Aikins, J.S., Kunz, J.C., Shortliffe, E.H. and Fallat, R.J. PUFF: an expert system for the interpretation of pulmonary function data. *Computers in Biomediccal Research* 1983, 16:199-208.

Altman, D.G. Statistics in medical journals: some recent trends. *Stat. Med.* 2000 19:3275-3289.

Altman, D.G. and Goodman, S. Transfer of technology from statistical journals to the biomedical literature: past trends and future predictions. *JAMA* 1994, 272:129-132

Altman, D.G. and Royston, P. What do we mean by validating a prognostic model ? *Stat. Med.* 2000 19:453-473.

Anderson CW, Stolz EA, Shamsunder S. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Biomed Eng* 1998, 45(3):277-286.

Andreassen, S., Benn, J.J., Hovorka, R., Olesen, K.G. and Carson, E.R. A probabilistic approach to glucose prediction and insulin dose adjustment - description of a metabolic model and pilot evaluation study. *Comp. Meth. Prog. in Biomed.* 1994, 41:153-166.

Astion, M.L, Wener, M.H., Thomas, R.G., Hunder, G.G. and Bloch, D.A. Overtraining in neural networks that interpret clinical data. *Clin. Chem.* 1993, 39(9): 1998-2004.

Bakken, I.J., Axelson, D., Kvistad, K.A., Brodtkorb, E., Muller, B., Aasly, J. and Gribbestad, I.S. Applications of neural network analyses to in vivo 1H magnetic resonance spectroscopy of epilepsy patients. *Epilepsy Res* 1999, 35(3):245-252.

Barach, P. and Small, P.D. Reporting and preventing medical mishaps: lessons from non-medical near-miss reporting systems. BMJ, 2000, 320:759-763.

Barnhill, S.D., Zhang, Z. and Madyastha, K.R. Evaluation of a new biochemical index for the estimation of bone demineralization using artificial intelligence. *Contemp Orthop* 1995, 30(4):315-318.

Baumgart-Schmitt R, Herrmann WM, Eilers R.On the use of neural network techniques to analyze sleep EEG data. Third communication: robustification of the classificator by applying an algorithm obtained from 9 different networks. *Neuropsychobiology* 1998, 37(1):49-58.

Baumgart-Schmitt R, Herrmann WM, Eilers R, Bes F. On the use of neural network techniques to analyse sleep EEG data. First communication: application of evolutionary and genetic algorithms to reduce the feature space and to develop classification rules. *Neuropsychobiology* 1997, 36(4):194-210.

Baxt, W.G. A neural-network trained to identify the presence of myocardial-infarction bases some decisions on clinical associations that differ from accepted clinical teaching. *Med. Dec. Making* 1994, 14(3):217-222.

Baxt, W.G. Improving the accuracy of an artificial neural network using multiple differently trained networks. *Neural Computation* 1992, 4(5):772-780.

Baxt, W.G. Use of an artificial neural network for data analysis in clinical decision making: the diagnosis of acute coronary occlusion. *Neural Computation* 1990, 2:480-489.

Baxt, W.G. Application of neural networks to clinical medicine. *Lancet* 1995, 346:1135-1138.

Baxt, W.G. Artificial neural network to identify acute myocardial infarction-Reply. *Lancet* 1996, 347:551.

Baxt, W.G. and Skora, J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. *Lancet* 1996, 347:12-15.

Baxt, W.G. and White, H. Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction. *Neural Computation* 1995, 7(3):624-638.

Biganzoli, E., Boracchi, P., Mariani, L. and Marubini, E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat. Med.* 1998, 17:1169-1186.

Bishop, C.M., Svensén, M., and Williams, C.K.I. GTM: the Generative Topographic Mapping. *Neural Computation* 1998, 10(1):215-234.

Bishop, C.M. Neural network for pattern recognition, Clarendon Press, Oxford 1995.

Blute, M.L., Bergstrahl, E.J., Partin, A.W., Walsh, OP.C., Kattan, M.W., Scardin, P.T., Montie, J.E., Pearson, J.D., Slezak, J.M. and Zincke, H. Validation of Partin tables for predicting pathological stage of clinically localized prostate cancer. *J. Urol.*, 2000, 164(5):1591-1595.

Boon, M.E. and Kok, L.P. Neural network processing can provide a means to catch errors that slip through human screening of Pap smears. *Diagn. Cytopathol.* 1993, 9:411-416.

Boon, M.E. and Kok, L.P. Histological validation of neural-network assisted cervical screening: a comparison with the conventional approach. *Cell Vision* 1995, 2:23-27.

Bounds, D.G., Lloyd, P.J. and Mathew, B. A comparison of nural networks and other pattern recognition approaches to the diagnosis of low back disorders. *Neural Networks* 1990: 583-591.

Brahams, D. and Wyatt, J. Decision-aids and the law. *Lancet* 1989, II:632-634.

Brennan, T.A., Leape, L.L., Laird, N.M., Hebert, L., Localio, A.R., Lawthers, A.G., *et al.* Incidence of adverse events and negligence in hospitalized patients. *N. Engl. J. Med.* 1991, 324:370-376.

Brown, S.F., Branford, A.J. and Moran, W. On the use of artificial neural networks for the analysis of survival data. *IEEE Trans. Neural Netw.* 1997, 8(5): 1072-1077.

Bryce ,T.J., Dewhirst, M.W., Floyd, C.E. Jr., Hars, V. and Brizel, D.M. Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck. *Int J Radiat Oncol Biol Phys* 1998, 41(2):339-345.

Bugliosi, R., Tribalto, M., Avvisati, G., Boccardoro, M., de Martinis, C., Friera, R., Mandelli, F., Pileri, A. and Papa, G. Classification of patients affected by multiple myeloma using neural network software. *Eur. J. Haematol.* 1994, 52(3):182-183.

Burke, H.B., Goodman, P.H., Rosen, D.B., Henson, DE., Weinstein, J.N., Harrell, Jr., F.E., Marks, J.R., Winchester, D.P. and Bostwick, D.G. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 1997, 79(4): 857-862.

Campbell, C. and Bennett, K.P. A linear programming approach to novelty detection. *In* Leen, T.K., Dietterich, T.G. and Tresp, V. (eds.), Advances in Neural Information Processing Systems 13, MIT Press, 2001.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A.L., Sandercock, P., Spiegelhalter, D. and Tryer, P. Framework for design and evaluation of complex interventions to improve health. *BMJ*, 2000, 321:694-696 and www.mrc.ac.uk/complex_packages.html

Cardoso, J.-F. Higher-order contrasts for independent components analysis. *Neural Computation* 1999, 11:157-192.

Caruana, R., Lawrence, S. and Giles, L. Overfitting in neural nets: backpropagation, conjugate gradient and early stopping. *In* Leen, T.K., Dietterich, T.G. and Tresp, V. (eds.), Advances in Neural Information Processing Systems 13, MIT Press, 2001.

Chan, H.P., Sahiner, B., Wagner, R.F. and Petrick, N. Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. *Med. Phys.* 1999, 26(1):2654-2668.

Chang GC, Luh JJ, Liao GD, Lai JS, Cheng CK, Kuo BL, Kuo TS. A neuro-control system for the knee joint position control with quadriceps stimulation. *IEEE Rehabil Eng* 1997, 5(1):2-11.

Chang R, Guan L, Burne JA. An automated form of video image analysis applied to classification of movement disorders. *Disabil Rehabil* 2000, 10-20;22(1-2):97-108.
Chen, H.Y., Chen, T.C., Min, D.I., Fischer, G.W. and Wu, Y.M. Prediction of tracolimus blood levels by using the neural network with genetic algorithm in liver transplantation patients. *Ther Drug Monit*. 1999, 21(1):50-56.
Chen, J.D., Lin, Z. and McCallum, R.W. Noninvasive feature-based detection of delayed gastric emptying in humans using neural networks. *IEEE Biomed. Eng.* 2000, 47(3):409-412.
Cherkassky, V. and Muller, F. Learning from data – concepts, theory and methods. John Wiley, New York 1998.
Civetta, J.M., Hudson-Civetta, J.A. and Nelson, L.D. Evaluation of APACHE II for cost containment and quality assurance. *Ann. Surg.* 1990, 212(3): 266-274.
Civetta, J.M., Hudson-Civetta, J.A., Kirton, O., Aragon, C. and Salas, C. Further appraisal of APACHE II limitations and potential. *Surg. Gynecol. Obstet.* 1992, 175(3): 195-203.
Coiera, E. Guide to medical informatics, the internet and telemedicine. London: Chapman & Hall Medical 1997.
Collett, D. Modelling survival data in medical research. Chapman and Hall, London, 1994:56-66.
Concato, J., Feinstein, A.R. and Holford, T.R. The risk of determining risk with multivariable models. *Ann Intern Med* 1993, 118:201-210.
Cross, S.S., Stephenson, T.J., Mohammed, T. and Harrison, R.F. Validation of a decision support system for the cytodiagnosis of fine needle aspirates of the breast using a prospectively collected dataset from multiple observers in a working clinical environment. *Cytopathology*, 2000, 11(6):503-512.
Cross, S.S, Harrison, R.F. and Lee Kennedy, R. Introduction to neural networks. *Lancet* 1995, 346:1075-1079.
Cuzick, J. and Sasieni, P. Cervical screening in the United Kingdom. *Hong Kong Med. J.* 1999, 5(3), 269-271.
Davies, M. and Owen, K. 'Complex uncertain decisions: medical diagnosis', Case Study 10 in Expert System Opportunities from the DTI's Research Technology Initiative, HMSO 1990.
Davies, R.J.O., Bennet, L.S., Barbour, C.,Tarassenko, L. and Stradling, J.R. Second by second patterns in cortical electroencephalograph and systolic blood pressure during Cheyne Stokes. *European Respiratory Journal* 1999, 14:940-945.
De Dombal, F.T., Leaper, D.J., Tanisland, J.R., McCann, A.P. and Horrocks, J.C. Computer-aided diagnosis of acute abdominal pain, *BMJ* 1972, 2:9-13.
De Dombal, F.T., Leaper, D.J., Horrocks, J.C., Stanisland, J.R. and McCann, A.P. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians, *BMJ* 1974, 1376-380.
De Dombal, F.T. Computer based assistance for medical decision making. *Gastroenterology and Clin. Biol.*, 8 1984: 135-137.
De Dombal, F.T., de Baere, H., van Elk, P.J., Fingerhut, A., Henriques, J., Lavelle, S.M., Malizia, G., Ohmann, C., Pera, C., Sitter, H. and Tsfitsis, D. Objective medical decision making - acute abdominal pain. *In* Beneken, J.E.W. and Thevenin, V. (eds.) Advances in biomedical engineering; results of the 4th EC Medical and Health research Programme. Vol. 7 of studies in health technology and informatics. Burke, Va.: IOS Press 1993:65-87.
De Dombal, F.T. Computer-assisted diagnosis in Europe. *N. Engl. J. Med.* 1994, 331(18): 1238.
De Dombal, F.T., Clamp, S.E. and Wardle, K.S. Measuring surgical performance in acute abdominal pain: some reflections from international studies. *Europ. J. Surg.* 1997, 163(5):323-329.
De Laurentiis, M. and Ravdin, P.M. A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Letters* 1994, 77:127-138.
De Laurentiis, M. and Ravdin, P.M. Survival analysis of censored data: neural network analysis detection of complex interactions between variables. *Breast Canc. Res. Treat.*1994, 32:113-118.

De Sutter, J., Van de Wiele, C., D'Asseler, Y., De Bondt, P., De Backer, G., Rigo, P. and Dierckx, R. Automatic quantification of defect size using normal templates: a comparative clinical study of three commercially available algorithms. *Eur J Nucl Med*, 2000, 27(12):1827-1834.

Dombi, G.W., Nandi, P., Saxe, J.M., Ledgerwood, A.M. and Lucas, C.E. Prediction of rib fracture injury outcome by an artificial neural network. *J Trauma* 1995 Nov;39(5):915-921.

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

Dreiseitl, S., Ohno-Machado, L. And Vinterbo, S. Evaluating variable selection methods for diagnosis of myocardial infarction. *Proc. AMIA Symp.* 1999, Part 1-2:246-250.

Drew, P.J. and Monson, J.R.T. Artificial neural networks. *Surgery*, 2000, 127(1):3-11.

Dreyfus, H. and Dreyfus, S. Why expert systems do not exhibit expertise. *IEEE Expert* 1986:86-90.

Dybowski, R. and Gant, V. Artificial neural networks in pathology and medical laboratories. *Lancet* 1995, 346:1203-11207.

Efron, B. Logistic regression, survival analysis and the Kaplan-Meier curve. *J. Am. Stats. Assoc.*1988, 83: 414-425.

Efron, B. and Tibshirani, R. Improvements on cross-validation: the .632+bootstrap method. *J. Am. Stat. Assoc.* 1997, 92(438):548-560.

Ellenius, J., Groth, T. and Lindahl, B. Neural network analysis of biochemical markers for early assessment of acute myocardial infarction. *Stud. Health Technol. Inform.* 1997, 43 Pt A: 382-385.

Ennis, M., Hinton, G., Naylor, D., Revow, M. and Tibshirani, R. A comparison of statistical learning methods on the GUSTO database. *Statstics in Medicine* 1998, 177:2501-2508.

Espinosa, J.A. and Nolan, W. Reducing errors made by emergency physicians in interpreting radiographs: longitudinal study. BMJ, 2000, 320:737-740.

Ezquerra, N. and Pazos, A. Neural computing in medicine. *editorial inartificial Intelligence in Medicine* 1994, 6:355-357.

Faraggi, D., Simon, R., Yaskil, E. and Kramar, A. Bayesian neural network models for censored data. *Biometrica J.* 1997, 5:519-532.

Finne, P., Finne, R., Auvinen, A., Juusela, H., Aro, J., Maattanen, L., Hakama, M., Rannikko, S., Tammela, T.L. and Stenman, U. Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network. *Urology*, 2000, 56(3):418-422.

Fricker, J. Artificial neural networks improve diagnosis of acute myocardial infarction. *Lancet* 1997,350:935.

Gaetz M, Weinberg H, Rzempoluck E, Jantzen KJ. Neural network classifications and correlation analysis of EEG and MEG activity accompanying spontaneous reversals of the Necker cube. *Brain Res Cogn Brain Res* 1998, 6(4):335-346.

Gamito, E.J., Stone, N.N., Batuello, J.T. and Crawford, E.D. Use of artificial neural networks in the clinical staging of prostate cancer: implications for prostate brachytherapy. *Tech Urol*, 2000,6(2):60-63.

Georgiadis, D., Kaps, M., Siebler, M., Hill, M., Konig, M., Berg, J., Kahl, M., Zunker, P., Diehl, B. and Ringelstein, E.B. Variability of Doppler microembolic signal counts in patients with prosthetic cardiac valves. *Stroke* 1995, 26(3):439-443.

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

Goldman, L., Cook, E.F., Brand, D.A., Lee, T.H., Rouan, G.W., Weisberg, M.C. *et al.* A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *N Engl J Med* 1988, 318:797-803.

Golub, R., Cantu Jr., R. and Tan, M. The prediction of common bile duct stones using a neural network. *J. Am. Coll. Surg.* 1998, 187(6):584-590.
Goodenday, L.S., Cios, K.J. and Shin, I. Identifying coronary stenosis using an image-recognition neural network. *IEEE Eng Med Bio Mag* 1997, 16(5):139-144.
Goodey, R.D., Brickley, M.R., Hill, C.M. and Shepherd, J.P. controlled trial of three referral methods for patients with third molars. *Br Dent J* 2000,189(10):556-560.
Groves, D.J., Smye, S.W., Kinsey, S.E., Richards, S.M., Chessells, J.M., Eden, O.B. and Bailey, C.C. A comparison of Cox regression and neural networks for risk stratification in cases of acute lymphoblastic leukemia in children. *Neural Comp. Appl.* 1999, 8(3):257-264.
Grozinger M, Kogel P, Roschke J. Effects of Lorazepam on the automatic online evaluation of sleep EEG data in healthy volunteers. *Pharmacopsychiatry* 1998, 31(2):55-59.
Grus, F.H. and Augustin, A.J. Analysis of tear protein patterns by a neural network as a diagnostic tool for the detection of dry eyes. *Electrophoresis* 1999, 20(4-5):875-880.
Gunning, K. and Rowan, K.M. ABC of intensive care: outcome data and scoring systems. *BMJ* 1999, 319:241-244.
Gurgen, F.S., Sihmanoglu, M. and Varol, F.G. The assessment of LH surge for predicting ovulation time using clinical, hormonal, and ultrasonic indices in infertile women with an ensemble of neural networks. *Comput. Biol. Med.* 1995, 25(4): 405-413.
Guterman H, Nehmadi Y, Chistyakov A, Soustiel JF, Feinsod M. A comparison of neural network and Bayes recognition approaches in the evaluation of the brainstem trigeminal evoked potentials in multiple sclerosis. *Int J Biomed Comput* 1996, 43(3):203-213.
Hanley, J.A. and McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982, 143:29-36.
Hanson, C.W. and Marshall, B.E. Artificial intelligence applications in the intensive care unit. *Crit. Care Med.*, 2001, 29(2):427-435.
Hayashi, Y., Setiono, R. and Katsumi, Y. A comparison between two neural network rule extraction techniques of hepatobiliary disorders. *Art. Intel. Med.* 2000, 20,3:205-216
Haynes, R.B. Loose connection between peer-reviewed clinical journals and clinical practice. *Ann. Intern. Med.* 1990, 113:724-728.
Hedén, B., Öhlin, H., Rittner, R. And Edenbrandt, L. Acute myocardial infarction detected in the 12-lead ACG by artificial neural networks. *Circulation* 1997, 96(6):1798-1802.
Heinrich H, Dickhaus H, Rothenberger A, Heinrich V and Moll GH. Single-sweep analysis of event-related potentials by wavelet networks-methodological basis and clinical application. *IEEE Biomed Eng* 1999, 46(7):867-879.
Hilden, J. Prevalence-free utility-respecting summary indices of diagnostic power do not exist. *Stat. Med.* 2000 19:431-440.
Horace Mann, N. and Brown, M.D. Artificial intelligence in the diagnosis of low back pain. *Orhtop. Clinics of North Am.* 1991, 22(2): 303-314.
Horrocks, J.C., McCann, A.P., Staniland, J.R., Leaper, D.J. and de Dombal, F.T. Computer-aided diagnosis: description of an adaptable system, and operational experience with 2,034 cases, *BMJ* 1972, 2:5-9.
Horwitz B, McIntosh AR, Haxby JV, Furey M, Salerno JA, Schapiro MB, Rapoport SI and Grady CL. Network analysis of PET-mapped visual pathways in Alzheimer type dementia. *Neuroreport* 1995, 6(17):2287-2292.
Hunt, D.L., Haynes, R.B., Hanna, S.E. and Smith, K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes. *JAMA* 1998, 280(15):1339-1346.
Husmeier, D., Penny, W.D. and Roberts, S.J. An empirical evaluation of Bayesian sampling with hybrid Monte Carlo for training neural network classifiers. *Neural Neworks* 1999, 12:677-705
Jadad A.R. and Rennie D. The randomized controlled trial gets a middle-aged checkup. *JAMA* 1998, 279:319-320.

Jain, A.K., Duin, R.P.W. and Mao, J. Statistical pattern recognition: a review. *IEEE-PAMI*, 2000, 22(1): 4-37.

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

Josefson, D. Computers beat doctors in interpreting ECGs. *BMJ* 1997, 315:763-766.

Kassirer, J.P. A report card on computer-assisted diagnosis - the grade C. *N. Engl. J. Med.* 1994, 330:1824-1825.

Kemeny V, Droste DW, Hermes S, Nabavi DG, Schulte-Altedorneburg G, Siebler M, Ringelstein EB. Automatic embolus detection by a neural network. *Stroke* 1999, 30(4):807-810.

Kennedy, R.L., Harrison, R.F., Burton, A.M., Fraser, H.S., Hamer, W.G., McArthur, D., McAllum, R., and Steedman, D.J. An artificial neural network system for diagnosis of acute myocardial infarction (AMI) in the accident & emergency department: evaluation and comparison with serum myoglobin measurements. *Computer Methods and Programs in Biomedicine* **19**97**,** 52**:** 93-103.

Kiani, K., Snijders, C.J. and Gelsema, E.S. Computerized analysis of daily life motor activity for ambulatory monitoring. *Technol Health Care* 1997, 5(4):307-318.

Kimberley, B.P., Kimberley, B.M. and Roth, L. A neuralnetwork approach to the prediction of pure tone thresholds with distortion product emissions. *Ear Nose Throat J.* 1994, 73(11):812-3, 817-23.

Knaus, W.A., Draper, E.A., Wagner, D.P. and Zimmerman, J.E. APACHE II: a severity of disease classification system. *Crit. Care Med.* 1985,13:818-829.

Knaus, W.A., Wagner, E.A., Draper, J.E., Zimmerman, M., Bergner, P.G., Bastos, C.A., Sirio, D.J., Murphy, D.J., Lotring, and Damiano, A. The APCAHCE III prognostic system: risk prediction of hospital mortality for critically hospitalised adults. *Chest* 1991, 100:1619-1636.

Kohn, L.T., Corrigan, J.M. and Donaldson, M.S. eds. *To err is human*: *building a safer health system.* Washington D.C: National Academy Press 1999.

Kol S, Thaler I, Paz N, Shmueli O. Interpretation of nonstress tests by an artificial neural network. *Am J Obstet Gynecol* 1995, 172(5):1372-1379.

Koss, L.G. The Papanicolaou test for cervical cancer detection: a triumph and a tragedy. *JAMA* 1989, 261: 737-743.

Koss, L.G., Sherman, M.E., Cohen, M.B., Anes, A.R., Darragh, T.M., Lemos, L.B., McClellan, B.J., Rosenthal, D.L., Keyhani-Rofagha, S., Schreiber, K. and Valente, P.T. Significant reduction in the rate of false-negative cervical smears with neural network-based technology (PAPNET Testing System). *Hum. Pathol.* 1997, 28(10):1196-1203.

Koss, L.G. The application of PAPNET to diagnostic cytology, in Lisboa, P.J.G., Ifeachor, E.C. and Szczepaniak, P.S. (eds.) 'Artificial neural networks in biomedicine' Springer, London, 2000:51-67.

Kothari, R., Cualing, H. and Balachander, T. Neural network analysis of flow cytometry immunophenotype data. *IEEE Biomed. Eng.* 1996, 43(8):803-810.

Kulikowski, C.A. Artificial intelligence in medical consultation systems: a review. IEEE-Eng. in Med. and Biology Mag. 1988: 34-39.

Lachebruch, P.A. Some misues of discriminants analysis. *Methods of Information in Medicine* 1977, 16:255-258

Ledley, R.S. and Lusted, L.B. Reasoning foundations of medical diagnosis. *Science.* 1959, 130:9-21.

Lee, Y.Y.B., Huang, Y., El-Deredy, W., Lisboa, P.J.G., Arús, C. and Harris, P. 'Robust Methodology for the Discrimination of Brain Tumours from in vivo Magnetic Resonance Spectra' *IEE Proceedings SMT* 2000, 147(6): 309-314.

Leistritz L, Hoffmann K, Galicki M, Witte H. Identification of hemifield single trial PVEP on the basis of generalized dynamic neural network classifiers. *Clin Neurophysiol* 1999, 110(11):1978-

1986.
Lemeshow, S. and Le Gall, J.R. Modelling the severity of illness of ICU patients. A systems update. *JAMA* 1994, 272(13): 1049-1055.
Leon, M.A. and Lorini, F.L. Ventilation mode recognition using artificial neural networks. *Comput Biomed Res* 1997, 30(5):373-378.
Liang H, Lin Z, McCallum RW. Application of combined genetic algorithms with cascade correlation to diagnosis of delayed gastric emptying from electrogastrograms. *Med Eng Phys* 2000, 22(3):229-234.
Liestøl, K., Andersen P.K. and Andersen, U. Survival analysis and neural nets. *Stat. Med.* 1994, 13:1189-1200.
Lindahl, D., Toft, J., Hesse, B., Palmer, J., Ali, S., Lundin, A. and Edenbrandt, L. Scandinavian test of artificial neural network for classification of myocardial perfusion images. *Clin Physiol*, 2000, 20(4):253-261.
Lindahl, D., Lanke, J., Lundin, A., Palmer, J. and Edenbrandt, L. Improved classification of myocardial bull's-eye scintigram with a computer-based decision support system. *J. Nuc. Med.* 1999, 40(1):96-101.
Lisboa, P.J.G. Industrial use of safety-related artificial neural networks. HSE CR 327/2001, HMSO 2001 and www.hse.gov.uk/research/crr_pdf/2001/crr01327.pdf
Lisboa, P.J.G., Ifeachor, E.C. and Szczepaniak, P.S. (eds.) 'Artificial neural networks in biomedicine' Springer, 2000a.
Lisboa, P.J.G., Vellido, A. and Wong, H. Bias reduction in skewed binary classification with Bayesian neural networks. *Neural Networks* 2000b, 13:407-410.
Lucas, P.J.F. Model-based diagnosis in medicine. Editorial, *Art. Int. Med.* 1997, 10:201-208.
Lundin, M., Lundin, J., Burke, H.B., Toikkanen, S., Pylkkänen, L. and Joensuu, H. Artificial neural networks applied to survival prediction in breast cancer. *Oncology* 1999, 57:281-286.
Lunn, D.J., Thomas, A., Best, N.G. and Spiegelhalter, D.J. WinBUGS- a Bayesian modelling framework: concepts, structure and extensibility. *Statistics in Computing*, 2000, 10:321-333.
Mackay, D. J. C., Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks network-computation in neural systems, *Network*: *Computation in Neural Systems* 1995, 6:469-505.
MacKay, D.J.C. Bayesian interpolation. *Neural Computation* 1992, 4:415-447.
Mango, L.J. Computer-assisted cervical cancer screening using networks. *Cancer Lett.* 1994, 77:155-162.
Mango, L.J. Clinical validation of interactive cytologic rescreening: automating the search, not the interpretation. *Acta Cytologica* 1997, 41(1): 93-97.
Mango, L.J. and Valente, P.T. Neural-network-assisted analysis and microscopic rescreening in presumed negative cervical cytologic smears. A comparison. *Acta Cytol* 1998 Jan-Feb;42(1):227-232.
Mariani, L., Coradini, D., Biganzoli, E., Boracchi, P., Marubini, E., Pilotti, S., Salvadori, B., Silvestrini, R., Veronesi, U., Zucali, R. and Rilke, F. Prognostic factors for metachronus contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast Canc. Res. Treat.* 1997, 44:167-178.
McAdam, W.A.F., Brock, B.M., Armitage, T., Davenport, P., Chan, M. and de Dombal, F.T. Twelve year's experience of computer-aided diagnosis in a district general hospital. *Ann. Roy. Coll. Surg. Engl.* 1990, 72:140-146.
McGuire, W.L., Tandon, A.K., Allred, D.C., Chamness, G.C., Ravdin, P.M. and Clark, G.M. Treatment decisions in axillary node-negative breast cancer patients. *J Natl Cancer Inst Monogr* 1992, (11):173-180.
Michaels, E.K., Niederberger, C.S., Golden, R.M., Brown, B., Cho, L. and Hong, Y. Use of a neural network to predict stone growth after shock wave lithotripsy. *Urology* 1998, 51(2):335-338.

Michie, D., Spiegelhalter, D.J. and Taylor, C. (eds.) Machine learning, neural nets and statistical classification. Ellis-Horwood, Chichester 1994.
Miller, R.A., Pople, H.E. Jr. and Myers, J.D. INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. N. Engl. J. Med. 307: 468-676 1982.
Mitchell, H., Medley, G. and Giles, G. Cervical cancers diagnosed after negative results on cervical cytology: perspective in the 1980s. *BJM* 1990, 300:1622-1626. Medline
Modai, I., Israel, A., Mendel, S., Hines, E.L. and Weizman R. Neural network based on adaptive resonance theory as compared to experts in suggesting treatment for schizophrenic and unipolar depressed in-patients. *J Med Syst*, 20(6):403-412.
Montie, J.E. and Wei, J.T. Artificial neural networks for prostate carcinoma risk assessment: an overview. *Cancer*, 2000, 88(12):2655-2660.
Moreno, R., Apolone, G. and Reis Miranda, D. Evaluation of the uniformity of fit of general outcome prediction models. *Intensive Care Medicine* 1998: 40-47.
Murray, L.S., Teasdale, G.M., Murray, G.D., Jennett, B., Miller, J.D., Pickard, J.D., *et al* Does prediction of outcome alter patient management ? *Lancet* 1993, 341:1487-1491.
Naguib, R.N., Adams, A.E., Horne, C.H., Angus, B., Sherbet, G.V. and Lennard, T.W. The detection of nodal metastasis in breast cancer using neural network techniques. *Physiol Meas* 1996,17(4):297-303.
Neal, R.M. 'Bayesian learning for neural networks' Springer-Verlag, New York 1996.
Ohmann, C., Moustakis, V., Yang, Q. and Lang, K Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. *AIM* 1996, 8:23-36.
Ohmann, C., Franke, C. and Yang, Q. Clinical benefit of a diagnostic score for appendicitis: results of a prospective interventional study. German study group of acute abdominal pain. *Arch. Surg*. 1999, 134(9):993-996.
Ohno-Machado, L. A comparison of Cox proportional hazards and artificial neural network models for medical prognosis. *Comput. Biol. Med.* 1997, 27(1):55-65.
Ohno-Machado, L. and Rowland, T. Neural network applications in physical medicine and rehabilitation. *Am. J. Phys. Med. Rehabil.* 1999, 78(4):392-398.
O'Leary T.J., Tellado, M., Buckner, S.B., Ali, I.S., Stevens, A. and Ollayos, C.W. PAPNET-assisted rescreening of cervical smears: cost and accuracy compared with a 100% manual rescreening strategy. *JAMA* 1998,279(3):235-237.
Park, J.H., Kari, S., King, L.E. and Olsen, N.J. Analysis of 31P MR spectroscopy data using artificial neural networks for longitudinal evaluation of muscle diseases: dermatomyositis. *NMR Biomed* 1998, 11(4-5):245-256.
Patterson, P. and Draper, S. A neural net representation of experienced and nonexperienced users during manual wheelchair propulsion. *J Rehabil Res Dev* 1998, 35(1):43-51.
Pesonen, E. Is neural network better than statistical methods in diagnosis of acute appendicitis ? *Stud. Health Technol. Inform.* 1997, 43(A):377-381.
Polak, M.J., Zhou, S.H., Rautaharju, P.M., Armstrong, W.W. and Chaitman, B.R. Using automated analysis of the resting twelve-lead ECG to identify patients at risk if developing transient myocardial ischaemia – an application of an adaptive logic network. *Physiol. Meas.* 1997, 18(4):317-325.
Potthoff, O., Schwefel, D., Rothemund, M., Engelbrecht, R. and van Eimeren, W. Expert systems in medicine. *Int. J. Technology Assessment* **4**:121-133 1988.
Prismatic Project Management Team. Assessment of automated primary screening on PAPNET of cervical smears in the PRISMATIC trial. *Lancet* 1999, 353(9162):1381-1385; *Erratum in* 353(9169):2078.
Radensky, P.W. and Mango, L.J. Interactive neural network-assisted screening: an economic assessment. *Acta Cytol*. 1998, 42:246-252.

Randolph, A.G., Haynes, R.B., Wyatt, J.C., Cook, D.J., Guyatt, G.H. and the Evidence Based Medicine Working Group. How to use an article evaluating the clinical impact of a computer-based clinical decision support system. *J.A.M.A.* 1999, 282(1):67-74.

Ravdin, P.M., Clark, G.M., Hilsenbeck, G., Owens, M.A., Vendely, P., Pandian, M.R. and McGuire, W.L. A demonstration that breast cancer recurrence can be predicted by neural network analysis. *Breast Canc. Res. and Treat.*1992, 21:47-53.

Reason, J. Human error: models and management. *BMJ*, 2000, 320:768-770.

Reggia, J.A. Neural computation in medicine. *Artificial Intelligence in Medicine* 1993, 5:143-157.

Riess, J. and Abbas, J.J. Adaptive neural network control of cyclic movements using functional neuromuscular stimulation. *IEEE Rehabil Eng* 2000, 8(1):42-52.

Ripley, R.M., Harris, A.L. and Tarassenko, L. Neural network models for breast cancer prognosis. *Neural Comput. Appl.* 1998, 7:367-375.

Ripley, B.D. Pattern recognition and neural networks. Cambridge University Press, Cambridge 1996.

Rogers, J., Jain, N.L. and Hayes, G.M. Evaluation of an implementation of PRODIGY Phase Two. Symposium of the *American Medical Informatics Association* 1999.

Rosenthal, D.L., Mango, L.J. Acosta, D. and Peters, R.K. "Negative" Pap smears preceding carcinoma of the cervix: rescreening with the PAPNET system. *Am. J. Clin. Pathol.* 1993, 100:331.

Rowan, K.M., Kerr, J.H., Major, E., McPherson, K., Short, A. and Vessey, M.P. Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. *Crit. Care Med.* 1994, 22:1392-1401.

Rutenberg, M.R. Neural network based automated cytological specimen classification system and method, United States Patent 4,965,725; 1990.

Savelberg HH and de Lange AL. Assessment of the horizontal,fore-aft component of the ground reaction force from insole pressure patterns by using artificial neural networks. *Clin Biomech (Bristol, Avon)* 1999, 14(8):585-592.

Schechter, C.B. Cost-effectiveness of rescreening conventionally prepared cervical smears by PAPNET testing. *Acta Cytol.* 1996, 40: 1272-1282.

Schmoor, C. and Schumacher, M. Effects of covariate omission and categorization when analysing randomised trials with the Cox model. *Stat. Med.* 1997, 15:225-137.

Schwartz, S., Wiles, J., Gough, I. and Phillips, S. Connectionist, rule-based and Bayesian desicion aids: an empirical comparison. In Hand, D.J. Artificial intelligence frontiers in statistics: AI and stats III, Chapman and Hall, London 1993:264-278.

Schwartz, W.B. Medicine and the computer: the promise and problems of change. *New Engl. J. Med.* 283 1970: 1257-1264.

Schwartz, W.B., Patil, R.S. and Szolovits, P. Sounding board: artificial intelligence - where do we stand ? N. Engl. J. Med. 1987, 316;685-688.

Schwartzer, G., Vach, W. and Schumacher, M. On the misuses of artificial neural networks for pronostic and diagnostic classification in oncology. *Stat. Med.*, 2000 19:541-551

Selker, H.P., Griffith, J.L., Patil, S., Long, W.J. and D'Agostino, R.B. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. *J Investig Med* 1995 Oct;43(5):468-476.

Shepherd, J.A. Computer-aided diagnosis of acute abdominal pain. Letter to the *BMJ* 1972, 3:347-348.

Sherman, M.E. and Kelly, D. High-grade squamous intraepithelial lesions and invasive carcinoma following the report of three negative Papanicolau smears: screening failures or rapid progression. *Mod. Pathol.* 1992, 5:337-342.

Sherman, M.E., Schiffman, M.H., Mango, L.J., Kelly, D., Acosta, D., Cason, Z., Elgert, P., Zaleski, S., Scott, D.R., Kurman, R.J., Stoler, M. and Lorincz, A.T. Evaluation of PAPNET testing as an ancillary tool to clarify the status of the "atypical" cervic al smear. *Mod Pathol* 1997, 10(6):564-571.

Shi, L.M., Fan, Y., Lee, J.K., Waltham, M., Andrews, D.T., Scherf, U., Paull, K.D. and Weinstein, J.N. Mining and visualizing large anticancer drug discovery databases. *Journal of Chemical Information and Computer Sciences*, 2000, 40(2):367-379.

Shortliffe, E.H. Computer-base medical consultations:MYCIN. New York: Elsevier 1976.

Shortliffe, E.H. Clinical decision-support systems. *In Shortliffe, E.H., Perreault, L.E., Wiederhold, G. and Fagan, L.M.* Medical informatics - computer applications in health care, Addison-Wesley, Reading, M.A. 1990.

Shortliffe, E.H. The adolescence of AI in medicine: will the field come of age in the '90s ? *Artif. Intell. Med.* 1993, 5:93-106.

Si, Y., Gotman, J., Pasupathy, A., Flanagan, D., Rosenblatt, B. and Gottesman, R. An expert system for EEG monitoring in the pediatric intensive care unit. *Electroencephalogr Clin Neurophysiol* 1998 Jun;106(6):488-500.

Simpson, H.W., McArdle, C., Pauson, A.W., Hume, P., Turkes, A. and Griffiths, K. A non-invasive test for the pre-cancerous breast. *Eur J Cancer* 1995, 31A(11):1768-1772.

Simpson, R.C. and Levine, S.P. Automatic adaptation in the NavChair Assistive Wheelchair Navigation System. *IEEE Rehabil. Eng.* 1999, 7(4):452-463.

Smith, B.P., Ward, R.A. and Brier, M.E. Prediction of anticoagulation during hemodialysis by population kinetics and an artificial neural network. *Artif Organs* 1998, 22(9):731-739.

Smith, J.H., Graham, J. and Taylor, R.J. The application of an artificial neural network to Doppler ultrasound waveforms for the classification of arterial disease. *Int J Clin Monit Comput* 1996, 13(2):85-91.

Sonke, G.S., Heskes, T., Verbeek, A.L., de la Rosette, J.J. and Kiemeney, L.A. Prediction of bladder outlet obstruction in men with lower urinary tract symptoms using artificial neural networks. *J Urol* 2000,163(1):300-305.

Spiegelhalter, D.J., Myles, J.J., Jones, D.R. and Abrams, K.R. An introduction to Bayesian methods in health technology. *BMJ* 1999, 319:508-512.

Stamey, T.A, Barnhill, S.D., Zhang, Z., Yemoto, C.M., Zhang, H. and Madyastha, K.R. Comparison of a neural network with high sensitivity and specificity to free/total serum PSA for diagnosing prostate cancer in men with a PSA < 4.0 ng/mL. *Mono. Urol.* 1998 19(2):21-32.

Stock, A., Rogers, M.S., Li, A. and Chang, A.M. Use of the neural network for hypothesis generation in fetal surveillance. *Baillieres Clin Obstet Gynaecol.* 1994, 8(3):533-548.

Szabo Z, Kao PF, Mathews WB, Ravert HT, Musachio JL, Scheffel U, Dannals RF. Positron emission tomography of 5-HT reuptake sites in the human brain with C-11 McN5652 extraction of characteristic images by artificial neural network analysis. *Behav Brain Res* 1996, 73(1-2):221-224.

Szczepaniak, P.S., Lisboa, P.J.G., and Kacprzyk, J. (eds.) 'Fuzzy systems in biomedicine' Springer-Verlag, Berlin, 2000.

Tafeit E, Moller R, Sudi K, Reibnegger G. The determination of three subcutaneous adipose tissue compartments in non-insulin-dependent diabetes mellitus women with artificial neural networks and factor analysis. *Artif Intell Med* 1999, 17(2):181-193.

Taktak, A.F., Simpson, S., Patel, S. and Meyer, G. Neural network analysis of oxygenation signals in infants during sleep. *Physiol Meas* 2000, 21(3):N11-22.

Thornhill, S., Teasdale, G.M., Murray, G.D., McEwen, J., Roy, C.W. and Penny, K.I. Disability in young people and adults one year after head injury: prospective cohort study. *BMJ*, 2000, 320: 1631-1635.

Tibshirani, R. and Knight, K. The covariance inflation criterion for adaptive model selection. *J. R. Stat. Soc. B* 1999, 61(3):529-546.

Tibshirani, R. A comparison of some error estimates for neural network models. *Neural Computation* 8 1996:152-163.
Vriesema, J.L., van der Poel, H.G., Debruyne, F.M., Schalken, J.A., Kok, L.P. and Boon, M.E. Neural network –based digitizaed cell image diagnosis of bladder wash cytology. *Diagn. Cytopathol.*, 2000, 23(3):171-179.
Walker, A.J., Cross, S.S. and Harrison, R.F. Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. *Lancet* 1999, 354:1518-1521.
Weiner, M.G. and Pifer, E. Computerized decision support and the quality of care. *Managed Care* 2000, 9(5):41-42, 444-6, 48-51.
Weingart, S.N., Wilson, R. McL., Gibberd, R.W. and Harrison, B. Epidemiology of medical error. BMJ, 2000, 320: 747-777.
Wilson, R.M., Runciman, W.B., Gibberd, Harrison, B.T., Newby, L. and Hamilton, J.D. The quality in Australian healthcare study. *Med. J. Aust* 1995, 163:458-471.
Winterer G, Ziller M, Kloppel B, Heinz A, Schmidt LG, Herrmann WM. Analysis of quantitative EEG with artificial neural networks and discriminants analysis- a methodological comparison. *Neuropsychobiology* 1998, 37(1):41-48.
Wong, L.S.S. and Young, J.D. A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks. *Anaesthesia* 1999, 54:1048-1054.
Wu, W.L. and Su, F.C. Potential of the back propagation neural network in the assessment of gait patterns in ankle arthrodesis. *Clin Biomech (Bristol, Avon)* 2000, 15(2):143-145.
Wyatt, J.C. Nervous about artificial neural networks ? *Lancet* 1995, 346:1175-1177.
Wyatt, J.C. and Altman, D.G. Commentary: prognostic models; clinically useful or quickly forgotten ? *BMJ* 1995, 311:1539-1541.
Zernikow, B., Holtmannspotter, K., Michel, E., Hornschuh, F., Groote, K. and Hennecke, K.H. Predicting length-of-stay in preterm neonates. *Eur. J. Pediatr.* 1999,158(1):59-62.
Zernikow, B., Holtmannspoetter, K., Michel, E., Theilhaber, M., Pielemeier, W. and Hennecke, K.H. Artificial neural network for predicting intracranial haemorrhage in preterm neonates. *Acta Paediatr.* 1998, 87(9):969-975.

Figure 1. Continuum of evidence, adapted from a generic model for complex healthcare interventions (Campbell *et al*, 2000).

Table 1. PUBMED entries involving neural networks listed under Randomised Controlled Trials (RCT).

Table 2. PUBMED entries involving neural networks listed under Clinical Trials (CT).

Table 3. RCT and CT with neural networks applied to prostatic, cervical and breast cancer.

Table 4. RCT and CT with neural networks in oncology.

Table 5. RCT and CT with neural networks in critical care.

Table 6. RCT and CT with neural networks in cardiology

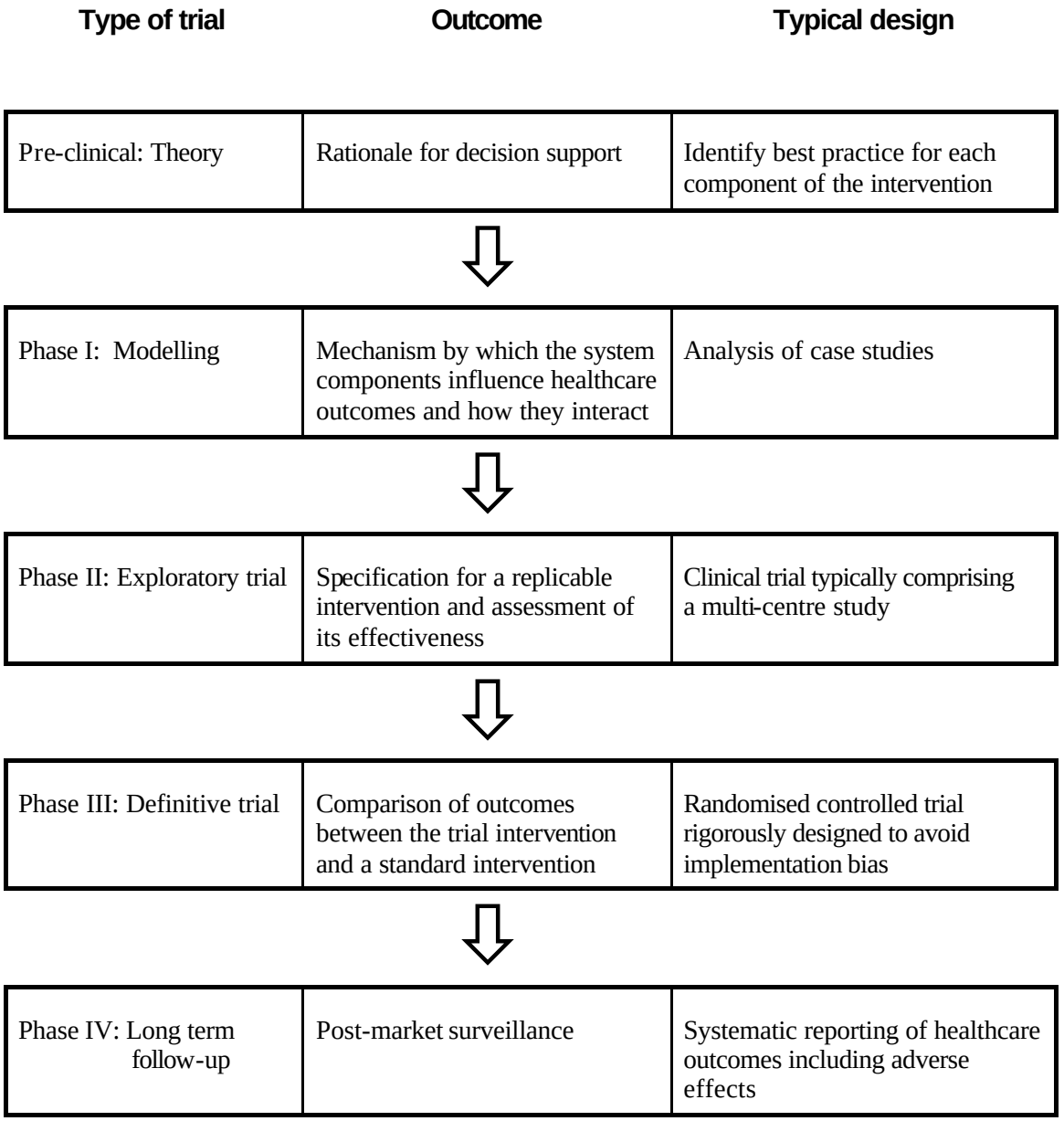
Type of trial
Outcome
Typical design
Pre-clinical: Theory
Rationale for decision support
Identify best practice for each component of the intervention
Phase I: Modelling
Mechanism by which the system components influence healthcare outcomes and how they interact
Analysis of case studies
Phase II: Exploratory trial
Specification for a replicable intervention and assessment of its effectiveness
Clinical trial typically comprising a multi-centre study
Phase III: Definitive trial
Comparison of outcomes between the trial intervention and a standard intervention
Randomised controlled trial rigorously designed to avoid implementation bias
Phase IV: Long term follow-up
Post-market surveillance
Systematic reporting of healthcare outcomes including adverse effects

Figure 1.

Table 1.

	Oncology	**Critical care**	**Cardiology**	**Other**
Diagnosis and staging	Prostatic cancer: Gamito *et al*, 2000 Finne *et al*, 2000 Cervical cancer: Doornewaard et al, 1999 Breast cancer: Naguib *et al*, 1996 Acute leukemia: Kothari *et al*, 1996	Intracranial haemorrhage in neonates: Zernikow *et al*, 1998	AMI: Ellenius *et al*, 1997 Baxt and White, 1995	Appendicitis: Pesonen, 1997
Outcome prediction	Response to therapy in head & neck cancer: Bryce *et al*, 1998 Recurrence of breast cancer in axillary node-negative patients: McGuire *et al*, 1992	Length-of-stay in preterm neonates: Zernikow *et al*, 1999		Tracolimus blood levels: Chen *et al*, 1999 Effect of treatment in schizophrenia and depression: Modai *et al*, 1996 Rib fracture injury: Dombi *et al*, 1995
Radiology	MRI of osteosarcoma: Glass and Reddick, 1998		Perfusion scintigraphy for detection of coronary stenosis: Goodenday *et al*, 1997 Doppler microembolic signal counts in patients with prosthetic heart valves: Georgiadis *et al*, 1995	
Physiological monitoring				Fetal surveillance during labour from fetal ECG: Stock *et al*, 1994

Table 2.

Clinical function	Oncology	Critical care	Cardiology	Neurology	Other
Diagnosis and staging	Cervical cancer: Prismatic team , 1999 Mango *et al*, 1998 Sherman *et al*, 1997 Pre-cancerous breast: Simpson *et al*, 1995		Transient ischaemia: Polak *et al*, 1997 Acute ischaemia: Selker *et al*, 1995	Embolus detection in stroke: Kemeny *et al*, 1999 Spontaneous EEG: Anderson *et al*, 1998 Sleep EEG: Baumgart-Schmitt *et al*, 1998 *et al*, 1997 Quantitative EEG: Winterer *et al*, 1998 Ventilation mode recognition: Leon and Lorini, 1997	Referral methods for patients with third molars: Goodey *et al*, 2000 Bladder outlet obstruction: Sonke *et al*, 2000 Tear protein patterns: Grus and Augustin , 1999 Haemodialysis: Smith *et al*, 1998 Ovulation time: Gurgen *et al*, 1995 Pure tone thresholds: Kimberley *et al*, 1994
Outcome prediction	Multiple myeloma: Bugliosi *et al*, 1994				Stone growth after lithotripsy: Michaels *et al*, 1998
Radiology			Myocardial perfusion images: Lindahl *et al*, 2000 Detection of stenoses from Doppler ultrasound waveforms: Smith *et al*, 1996	MRS of epilepsy: Bakken *et al*, 1999 PET of 5-HT reuptake sites: Szabo *et al*, 1996. PET in Alzheimer's: Horwitz *et al*, 1995	MRS of muscle: Park *et al*, 1998
Physiological monitoring		EEG in Pediatrics: Si *et al*, 1998		Single trial PVEP: Liestritz *et al*, 1999 Heinrich *et al*, 1999 Correlation of EEG and MEG: Gaetz *et al*, 1998 Lorazepan and sleep EEG: Grozinger *et al*, 1998 Evoked potentials in multiple-sclerosis: Guterman *et al*, 1996	Oxygenation in infants: Taktak *et al*, 2000 EGG of gastric empting: Liang *et al*, 2000 Chen *et al*, 2000 Subcutaneous adipose tissue: Tafeit *et al*, 1999 Nonstress tests in obstetrics: Kol *et al*, 1995 Bone dimeneralization: Barnhill *et al*, 1995
Other	Gait patterns: Chang *et al*, 2000, Wu and Su, 2000 Ground reaction force: Savelberg *et al*, 1999 Wheelchair navigation: Simpson and Levince, 1999 Wheelchair propulsion: Patterson and Draper, 1998 Daily motor activity: Kiani *et al*, 1997 FES: Riess and Abbas, 2000, Abbas and Triolo, 1997, Chang *et al*, 1997				

Table 3.

Reference	No. of subjects	Clinical function	Performamce assessment	Conclusions
Prostatic cancer				
Gamito *et al*, 2000	4,133	Prediction of risk of lymph node spread (LNS) from age, race, PSA, PSA velocity, Gleason sum and TNM	External validation (n=660)	98% accuracy in detection of low risk of LNS with a MLP
Gamito *et al*, 2000	409	Predicting capsular penetration (CP) with the same explanatory variables as above	Train/test	84% accuracy in detection of CP with a MLP
Finne *et al*, 2000	656	Elimination of false-positive PSA results by combining total PSA, proportion of free PSA, digital rectal examination and prostate volume	Leave-one-out	At clinically relevant sensitivities MLP and LogR reduce the number of false-positives significantly better than the proportion of free PSA
Cervical cancer				
Prismatic team, 1999	NNA	Assessment as a primary screening tool for categorization of cervical smears adequate for reporting as negative, mild, moderate or severe dyskaryosis, invasion, glandular neopplasia and borderline nuclear changes	External validation (n=21,700)	89.9% agreement across all classes was found between PAPNET and conventional primary screening, with similar sensitivity (82 cf. 83%), with PAPNET having improved specificity (77 cf. 42%) and faster processing (3.9 min. cf. 10.4 min)
Doornewaard *et al*, 1999	NNA	Assessment as a primary screening tool for the early detection of cervical dysplasia	External validation (n=6,063)	PAPNET testing has similar diagnostic value to conventional screening of Pap smears, with AUROC 95% CIs of 78-82% for control and 77-81% for PAPNET
Mango *et al*, 1998	NNA	Comparison of yield in re-screening of node-negative PAP smears between NNA and conventional unassisted cytology	External validation (n=10,000)	PAPNET returned a yield of 6.2% versus 0.6% for manual re-screening
Sherman *et al*, 1997	NNA	Evaluation of an ancillary tool to clarify the status of atypical smears with borderline abnormalities, comparing the results of 5 trained cytologists	External validation (n=200)	Consensus PAPNET results of abnormal were predictive of abnormal histological findings at follow-up
Breast cancer				
Naguib *et al*, 1996	81	Prediction of the lymph node involvement with surrogate measurements of the primary tumour, by combining SOM and MLP layers	Train/test/ validation	Neural networks are sensitive for predicting lymph node positive patients
Simpson *et al*, 1995	91	Identification of pre-cancerous from normal breast based on thermal profiles combined with progestorone and steroid measurements	Train/test	Sensitivity and specificity above 90% for LDA and MLP for aged matched patients
McGuire *et al*, 1992	199	Prediction of the the likelihood of a relapse within 5 years in axillary node-negative patients	Train/test	The MLP was more specific than conventional analysis for the identification of low-risk patients

Table 4.

Reference	No. of subjects	Clinical function	Performamce assessment	Conclusions
Head and neck				
Bryce *et al*, 1998	95	Survival prediction following treatment for squamous cell carcinoma	Cross-validation	MLP modelled uncensored survival better than LogR, with AUROC 95% CIs of 78% ± 5% and 67% ± 5% respectively (p=0.07) and better than clinical staging alone (60% ± 7%, p<0.02)
Osteosarcoma				
Glass and Reddick, 1998	43	Assessment of percentage of necrosis from dynamic contrast enhanced MRI in a two-step process involving segmentation with SOM and quantification with MLP	Train/test/ validation	The predicted percentage of necrosis and histopahological analysis correlated with a Spearman rank coefficient of 0.617 (p<0.001)
Acute leukemia				
Kothari *et al*, 1996	170	Categorisation into subcategories based on lineage and differentiation in antigen expression with 28 available covariates	Train/test	MLP regularised with weight decay generalised with 10% misclassification error from lineage or differentiation
Multiple myeloma				
Bugliosi *et al*, 1994	172	Prediction of survival from patient characteristics at onset at response to induction therapy.	Train/test	Although test performance is perfect, this study is recognised to be preliminary

Table 5.

Reference	No. of subjects	Clinical function	Performamce assessment	Conclusions
Neonates				
Zernikow *et al,* 1999	2,144	Predicting length-of-stay in preterm neonates from 40 first-day-of-life items	Train/test/ validation	First-day-of-life data is predictive of length-of-stay of pre-term neonates with correlation CIs of 0.85-0.90 for MLR and 0.87-0.92 for MLP
Zernikow *et al,* 1998	890	Intracranial haemorrhage in neonates from admission data	Train/test	AUROC 0.935 for MLP and 0.884 for LogR. MLP more accurate than LogR (p=0.001) with also better sensitivity
Si *et al,* 1998	74	Warning system for the pediatric intensive care unit (PICU) about EEG abnormalities	Cross-validation	Expert system based on neural networks and fuzzy logic agreed with an expert for 45% EEG sections and predicted with 1 of 7 levels of abnormality in 91%
Stock *et al*, 1994	113	Prediction of umbilical artery pH from 13 fetal ECG features	Train/test	The predictions from a MLP correlated with measured pH significantly better than those form MLR

Table 6.

Reference	No. of subjects	Clinical function	Performamce assessment	Conclusions
AMI				
Ellenius *et al,* 1997	88	Early diagnosis or exclusion of AMI and staging of infarct from biochemical markers	Train/test	MLP approach could provide useful support for assessment of patients with suspectd AMI
Baxt and White, 1995	706	Detection of AMI in emergency departments, from *et al,* 19 variables representing patient history, clinical findings and ST-T measurements taken from the ECG	Bias-corrected sensitivity analysis using the bootstrap	Potential for large bias is present in direct measurements of input effects, and a methodology is proposed to remove this bias
Ischaemia				
Polak *et al,* 1997	1,367	Prediction of transient ischaemia during ambulatory Holter monitoring, from a resting 12-lead ECG. Univariate t-tests were used to inform model selection	Train/test	LDA and adaptive logic networks were superior to the MLP to predict the likelihood for the occurrence of ischaemic episodes
Selker *et al,* 1995	3,453	Clinical indicators available within 10 minutes of emergency department care were used to predict AMI and unstable angina pectoris, in a real-time clinical setting	External validation (n=2,320)	Limiting the inputs to 8 readily available variables, AUROCs for LogR, CART and MLP were 0.887, 0.858 and 0.902, respectively. Each is a clinically useful predictor of clinical outcome
Radiology				
Lindahl *et al,* 2000	135	Detection of coronary disease from myocardial perfusion scintigrams	External validation (n=68)	At a clinically relevant specificity the sensitivity of a MLP was significantly better than one of six clinical criteria and two CEqual-based criteria
Goodenday *et al,* 1997	42	Diagnosis of coronary stenosis from radionuclide myocardial perfusion scintigraphy, using a hierarchical unsupervised image-recognition neural network modelled on the neocognitron	Train/test	Identification of coronary artery stenosis from unprocessed clinical images was good and compared favourably with alternative computer-based methods, but some difficulties were encountered with rotation and scale invariance
Smith *et al,* 1996	219	Detection of stenosis at the site of the common femoral artery, distinguing waveforms from proximal, distal and multi-segmented sites	Cross-validation	Separation of stenosis sufferers from healthy controls was possible and more effective with an MLP than with a Bayesian classifier.
Georgiadis *et al,* 1995	73	Detection of microembolic signals in patients with prosthetic heart valve by means of a single transcranial Doppler ultra-sound 30 minute session	Train/test	There were no significant differences beween signal counts detected by a MLP or by trained observers from 3 centres