Pattern Recognition in Physiological Time Series Data Using Bayesian Neural Networks

Timothy Paul Howells

Ph.D.
University of Edinburgh
2002

I declare that this thesis is entirely my own work

Tim Howells

7 February 2002

Abstract

Neural networks have been used successfully in many important applications. Speech recognition, optical character recognition and image processing are examples of areas where neural networks have become one of the standard solutions to difficult problems in automatic pattern recognition. This success has generated interest in the scientific community among researchers looking for more powerful tools than the standard parametric statistical models for the analysis of complex datasets. Progress in this area has been more problematic. The behavior of neural networks can be notoriously difficult to understand or interpret. Although their asymptotic properties have been well understood for a long time, model validation has required large training, validation and test data sets, which is seldom feasible in the context of scientific research. In recent years this has started to change. Progress is being made towards understanding the statistical bases of neural network training and performance from a number of different perspectives (Bishop, 1995). One important line of research in this area is the application of Bayesian techniques to network learning (Neal, 1996).

This thesis describes the application of these Bayesian techniques to the analysis of a large database of physiological time series data collected during the management of patients following traumatic brain injury at the Western General Hospital in Edinburgh. The study can be divided into three main sections:

- *Model validation using simulated data:* Techniques are developed that show that under certain conditions the distribution of network outputs generated by these Bayesian neural networks correctly models the desired conditional probability density functions for a wide range of simple problems for which exact solutions can be derived. This provides the basis for using these models in a scientific context.

- *Model validation using real data.* Statistical prognostic modelling for head injured patients is well advanced using simple demographic and clinical features. The Bayesian techniques developed in the previous section are applied to this problem, and the results are compared to those obtained using standard statistical techniques.

- *Application of these models to physiological data.* The models are now applied to the full database and used to interpret the data and provide new insight into the risk factors for head injured patients in intensive care.

References

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition,* Oxford University Press

Neal, R.M., 1996. *Bayesian Learning for Neural Networks,* Springer. Lecture Notes in Statistics 118

Table of Contents

Chapter 1: Introduction **1**
1.1 Neural networks and scientific research 1
1.2 Plan of the thesis 2
1.3 The medical application: Understanding the mechanisms of traumatic brain injury 4
1.4 The Edinburgh head injury database 5
1.5 Methodology 7
1.5.1 Using neural networks to support medical research 7
1.5.2 Bayesian Neural Networks 8
1.5.3 The hybrid Monte Carlo algorithm 12
1.5.4 Discussion of hybrid Monte Carlo 15
1.5.5 Adaptive regularization 17
1.5.6 Bayesian model selection 19
1.5.7 The application of Bayesian neural networks in a scientific context 21
1.6 Alternative technical approaches 22
1.6.1 Gaussian processes 22
1.6.2 Support Vector Machines 23
1.6.3 Neural networks with bootstrap 24
1.6.4 Optimization on choice of input features 25
1.6.5 Decision trees and bagging 26
1.6.6 General rule discovery and boosting 28
1.7 Statistical prognostic modeling and head injury research 29
1.8 Applications of artificial intelligence in intensive care 31
1.8.1 Clinical audit of patient management 32
1.8.2 Intelligent multiparameter alarms 32
1.8.3 Early warning systems 33
1.8.4 Decision support 33
1.8.5 Pattern recognition and machine learning 34
1.9 Summary 35

Chapter 2: Modeling two kinds of uncertainty **37**
2.1 A dilemma in data analysis 37
2.2 Bayesian inference and neural networks 39
2.3 Interpreting the outputs of individual networks 41
2.4 Interpreting the MCMC output distribution 42
2.5 Calibrating the prior distribution 43
2.6 Modeling Bernoulli trials 46
2.7 Interpreting the output distributions 49
2.8 The importance of tuning the prior 50
2.9 The noisy XOR problem 53
2.10 Discussion of the simulation results 56
2.11 Probability distributions over three outputs 57
2.12 Validating the three output model 58
2.13 Input rescaling 58
2.14 Selecting the prior 60
2.15 Generalization 61
2.16 Summary 64

Chapter 3: Prognostic models based on demographic data and simple clinical indicators **67**
3.1 Baseline models using data available on admission 67
3.2 The motor score model 68
3.3 Modeling the effect of age 71
3.4 Comparing models 72
3.5 Comparisons of some simple models 75
3.6 Comparison with previous work on another database 78
3.7 Other work on the Edinburgh database 80
3.8 Model performance with increased sample size and increased information per sample 82
3.9 Summary 85

Chapter 4: Feature extraction from physiological time series data **87**
4.1 The Edinburgh University Secondary Insult Grades 87
4.2 A study based on the Traumatic Coma Data Bank 89
4.3 Summary 91

Chapter 5: Raised intracranial pressure and related factors **93**
5.1 Intracranial pressure and cerebral perfusion pressure 93
5.2 Controlling for admission factors 102
5.3 The time course of ICP and CPP insults 105
5.4 Arterial hypotension 109
5.5 Pyrexia 114
5.6 Summary 117

Chapter 6: Clinical factors relating to cerebral oxygenation **121**
6.1 Hypoxia 121
6.2 Cerebral Hyperemia 126
6.3 Cerebral Oligemia 129
6.4 The time course of SvO2 insults 129
6.5 Summary 131

Chapter 7: Multivariate models combining physiological monitoring data and admission data **133**
7.1 Model Comparison 133
7.2 Summary 139

Chapter 8: Conclusions and future work **141**
8.1 Validating the use of Bayesian neural networks in scientific research 141
8.2 Clinical conclusions 143
8.3 Unresolved issues and a criticism 144
8.4 Future work 146

Acknowledgments **147**

Appendix 149
A.1 Parzen density estimation 149
A.1.1 Two output case 149
A.1.2 Three output case 150
A.2 Beta and Dirichlet functions 151
A.2.1 Two output case (Beta function) 152
A.2.2 Three output case (Dirichlet function) 152
A.3 Parameters for the simulations 153
A.3.1 Neural network model specifications 153
A.3.2 Monte Carlo Markov chain specifications 154

References **155**

Chapter 1

Introduction

1.1 Neural networks and scientific research

Neural Networks have been used successfully in many important applications. Speech recognition, optical character recognition and image processing are examples of areas in which neural networks have become one of the standard solutions to difficult problems in automatic pattern recognition. This success has generated interest in the scientific community among researchers looking for more powerful tools than the standard parametric statistical models for the analysis of complex datasets. Progress in this area has been more problematic. The behavior of neural networks can be notoriously difficult to understand or interpret. Although their asymptotic properties have been well understood for a long time, model validation has required large training, validation and test data sets. It is seldom feasible for the scientific researcher to collect such large quantities of data.

In recent years this has started to change. Progress is being made towards understanding the statistical bases of neural network training and performance from a number of different perspectives (Bishop, 1995). One important line of research in this area is the application of Bayesian techniques to network learning (Neal, 1996). This work has demonstrated principles by which model complexity can be automatically adapted on the basis of the available data and confidence regions can be assigned, taking into account model uncertainty. Unlike the parametric techniques typically employed in

medical research, these techniques make minimal assumptions about the nature of the training data.

The work described in this thesis applies Bayesian neural networks to a significant "real world" problem. This has required a thorough exploration of the practical and theoretical issues encountered when using these models in a scientific context. This is not a comparative study. I decided early on that because of the inherent importance of the subject matter, I would do an in depth study rather than several simple studies based on alternative technical approaches. I therefore can't claim that the use of Bayesian neural networks is the best approach possible, but I hope that I have shown that it is a good approach and that this study has advanced the state of knowledge in the application area.

1.2 Plan of the thesis

This thesis describes the application of these Bayesian techniques to the analysis of a large database of physiological time series data collected during the management of patients following traumatic brain injury at the Western General Hospital in Edinburgh. The study can be divided into three main sections:

- *Model validation using simulated data:* Techniques are developed that show that under certain conditions, the distribution of network outputs generated by these Bayesian neural networks correctly models the desired conditional probability density functions for a wide range of simple

problems for which exact solutions can be derived. This provides the basis for using these models in a scientific context. (Chapter 2)

- *Model validation using real data:* Statistical prognostic modeling for head injured patients is well advanced using simple demographic and clinical features. The Bayesian techniques developed in the previous section are applied to this problem, and the results are compared to those obtained using standard statistical techniques (Chapter 3).
- *Application of these models to physiological data:* The neural network models are now applied to the full database, and used to interpret the data and provide new insight into the risk factors for head injured patients in intensive care (Chapters 4 - 7).

The remainder of this chapter will discuss the medical application that is the subject of this thesis, and the methodology employed in analyzing this data. Section 1.3 will describe the medical problem being studied. Section 1.4 will describe the Edinburgh headinjury database. Section 1.5 concerns the implementation and interpretation of Bayesian neural networks. Alternative technical approaches will be described in section 1.6. Background on existing work in the field will be provided in section 1.7, and alternative applications of artificial intelligence in intensive care will be discussed in section 1.8.

1.3 *The medical application: Understanding the mechanisms of traumatic brain injury*

An issue that has stimulated research in the field of head injury treatment is the time course of the pathological processes that follow brain trauma. Researchers have long remarked on patients who "talk and die". Following brain injury a patient will sometimes recover his faculties to a very large extent and appear to be doing well only to then deteriorate and ultimately die as a result of the injury. This indicates that the damage sustained by the brain is not an immediate effect of the primary injury, but rather develops over a period of several days. This observation has been confirmed by several studies of neuronal and structural brain damage (see Miller 1992, for a review). The use of therapeutic agents to intervene in this process and protect the brain has been much investigated, but with limited success. The hope remains that our increasing understanding of cerebral hemodynamics and the "biochemical cascade" that occurs following brain trauma will ultimately lead to the development of techniques that will allow us to protect the brain during this critical period.

Due to the complexity of the cerebrospinal system and the demands of the neurosurgical environment there has been much interest in the development of new monitoring technology which may detect the causes of morbidity and brain damage in patients in neurointensive care. Transcranial Doppler devices have been used to measure cerebral blood flow velocity (Chan et al., 1992). The use of jugular bulb oxygenation measurements to estimate brain oxygen extraction has been investigated (Gopinath et al, 1994). EEG can be used to detect subclinical seizures (Vespa, et al., 1997). Intracerebral microdialysis is used to measure changes in brain metabolism

(Persson and Hillered, 1992). A device for continuous measurement of the compliance of the cerebrospinal system is now available (Piper et al. 1999).

This increasing interest in multimodality monitoring has in turn underlined the need for computers in neurointensive care for data acquisition, integration and analysis. This has led to the development of several research software systems dealing with various aspects of the problem. (Czosnika et al., 1994) describes a system focused on the use of signal processing techniques for early detection of acute episodes in the patients being monitored. A study of jugular bulb oxygen saturation (Gopinath et al, 1994) was enabled by the use of computerised monitoring. The next section will describe a system for multimodality computerised monitoring developed at Edinburgh University to support research in the management of patients following traumatic brain injury (Piper et al. 1991, Howells et al, 1995).

1.4 The Edinburgh Head Injury Database

The work described in this thesis was conducted in support of a study of pathophysiological factors following head injury initiated by the late professor Douglas Miller and funded by the Medical Research Council. This study introduced computerised monitoring into the intensive care unit at the Western General Hospital in Edinburgh. The computers are attached to patient monitors via their serial communication ports, and collect trended samples of the physiological data once per minute. The first patients included in this database were admitted in December of 1991. The most recent that are included in the study reported here were admitted in October 1998. Basic demographic and clinical data are available for 719 patients. Of these 243 were subject to

computerized monitoring. In general I have restricted myself to patients over the age of 14 who were classified as having a severe head injury. This limits the numbers to 286 patients altogether and 158 with computerized monitoring. For any particular model, further restrictions may be introduced due to missing data.

Earlier work on this database led to the development of the Edinburgh University Secondary Insult Grades, which define a set of adverse physiological events, such as episodes of hypotension and raised intracranial pressure which can be extracted from this detailed record and analysed. One study of secondary insults showed that computerised monitoring recorded physiological derangements that were missed on the nurses' chart (Corrie et al., 1993). Another study showed that the occurrence of "secondary insults" bore a statistical relationship to patient outcome (Jones et al, 1994). My role on this project has been to revise and extend the software so that it can be used as a clinical tool. The system developed is now being used in seven intensive care units in Britain, Italy, Switzerland and Sweden. In addition to head-injury, it has been used to support research in the management of patients following cardiac surgery, and stroke. The pattern recognition techniques described in this thesis are integrated into the software.

1.5 Methodology

1.5.1 Using neural networks to support medical research

This study is based on detailed physiological monitoring of patients in intensive care. Readings for parameters such as arterial blood pressure, temperature and intracranial pressure were automatically recorded on bedside computers once a minute. The duration of computerized monitoring for each patient varied from several hours to almost three weeks. I have applied Bayesian neural networks as prognostic models taking as input demographic data, clinical indicators, and features extracted from the physiological time series data. The networks are trained to predict outcome probabilities, for example the probability of survival given the input features for a patient. I have then used the behavior of these models to gain insight into the database, and the risk factors facing patients following brain trauma.

This methodology is contrary to conventional wisdom regarding the "black box" nature of neural networks. Ripley (1996) has stated that "neural networks have almost no explanatory power", while by contrast,

> "Linear regression has traditionally been taught from the viewpoint of explanation, which reflects its importance in that role in scientific and medical research." (Ripley, 1998)

Similarly, Dybowski and Weller (1999) have argued,

> "The complexity of neural networks does make it difficult to grasp how their output relates to input. Hart and Wyatt (1990) believe that this 'black box' aspect is a major obstacle to the acceptance of neural nets as part of medical decision support systems ... We think that a neural net (which can be regarded as a complex regression model) can be accepted in medicine with or without a detailed understanding of how it works – provided its predictive capability has been rigorously evaluated."

Contrary to the view that the value of neural networks is at best limited to making accurate predictions to support decision making, the study reported here has demonstrated that they *can* be used to gain insights into a complex medical data set, and that these insights can be translated into straightforward medical guidelines that can be supported by reference to the original data being modeled. These results differ from those obtained using linear logistic regression, and demonstrate the utility of these flexible, non-linear models in a scientific context. This work builds on current research into the application of Bayesian inference to the development of neural network models. In chapter two it will be shown that these techniques lead to neural network systems that produce an output distribution that models the distribution of the target values conditioned on the inputs. These distributions in output space quantify firstly, the uncertainty in prediction due to incomplete information in the input variables, and secondly, the uncertainty in that estimate due to the amount of available training data. This clear mathematical interpretation of results leads to systems that can be used quite naturally in scientific research, unlike the "black box" neural network implementations.

1.5.2 Bayesian neural networks

The Bayesian approach to neural network learning begins with the simple observation that a finite data set cannot tell us with probability one the exact values that the network weights should assume. The maximum likelihood approach to estimating the weights of a neural network ignores this inconvenient fact. If very large amounts of training data are available, this may be an appropriate, or even

necessary, modeling choice. However, the application described here requires a more principled approach. Using Bayes' theorem we can write down the probability density function for network weights (**w**) conditioned on the training data (D) as:

$$p(\mathbf{w}\,|D) = \frac{p(D|\mathbf{w})\,p(\mathbf{w})}{\int p(D|\mathbf{w})\,p(\mathbf{w})\,d\mathbf{w}}$$

For a given weight vector, the two terms that have to be calculated on the right hand side are:

- $p(D|\mathbf{w})$: The probability of the training data given a set of weights
- $p(\mathbf{w})$ The prior probability of the weights

In our application, a classification task, the first quantity can be calculated using the cross entropy, or multi-logistic error term, which has been shown to model this conditional probability (e.g. see Bishop, 1995 pg. 230 ff.). The second term, the prior probability for a weight vector, is generally chosen to be a circularly symmetric Gaussian with zero mean. This reflects the fact that large weights lead to unstable systems that do not generalize well to new data: hence the general preference for smaller weights. The use of this prior has been demonstrated to be equivalent to the use of weight decay with a sum squared term for the weight error (MacKay, 1992a), a technique which has been used effectively to regularize neural networks for many years. The choice of prior for the models used here will be discussed in detail in chapter two.

This leaves us with the problem of integrating over the posterior in weight space. One approach to this problem (MacKay, 1992a) is to find one or more local

maxima of the distribution using a standard technique for optimizing network weights, for example, conjugate gradients. Then the distributions around these peaks can be approximated as Gaussians using the Hessian matrix of the error term with respect to the weights computed at the local maxima. This technique has been used effectively in many applications. However, there are concerns about the quality of the Gaussian approximation, especially when the networks used have large numbers of parameters and in the presence of sparse data (Bishop, 1995). Classification tasks also pose a problem for this approach, since the nonlinear transformation into output space results in distributions that are far from Gaussian. This then requires a correction term on top of the other analytical approximations (MacKay, 1992c). All of these concerns regarding Gaussian approximation of the weight posterior will be prominent in this application.

A second approach to approximating the posterior on network weights is to generate a series of weight vectors from the posterior to build up a discrete approximation. A simple way to do this would be to use rejection sampling. That is, to generate samples from the prior, which can then be accepted or rejected according to their posterior probability as estimated using the error term. This would be prohibitively expensive computationally because the posterior is typically so sharply peaked around the local minima of the error that virtually all of the weight vectors generated in this way would be from regions of weight space having negligible probability, and would therefore be rejected. Improved sampling techniques can be implemented using Monte Carlo Markov chain (MCMC) techniques. These introduce dependencies between successive samples, so that weight space is explored more systematically. From a random starting point the chain converges

towards the posterior distribution. Once it is judged to have converged to the posterior, the process continues and samples from the posterior are generated to form the discrete approximation.

An example of a Monte Carlo Markov chain method is the Metropolis algorithm, in which a series of samples is generated through a series of short, random steps through weight space. From each point, candidate steps are generated from a "proposal distribution", and accepted with probability $\min(1, \frac{p(\mathbf{w}_{current})}{p(\mathbf{w}_{candidate})})$. The proposal distribution used to generate candidates must be symmetrical. That is, $p(\mathbf{w}_i|\mathbf{w}_j) = p(\mathbf{w}_j|\mathbf{w}_i)$. A reasonable choice would be a Gaussian centered on the current point. The width of the Gaussian should be sufficiently narrow that once a good region of weight space is entered, the rejection rate will be fairly low. The Metropolis algorithm would be a reasonable way to generate samples from the posterior distribution of weights for a neural network. However, recent work has demonstrated that there are much more efficient methods, which will be discussed below.

It's important to point out that the posterior on network weights is of little interest itself. The contribution of the Bayesian approach is that the distribution in weight space produces a distribution in *output* space. This allows us to study the relationship between the input and target data in greater detail than has previously been possible, especially in the presence of sparse and noisy data. Much of the work in this thesis has involved visualization and analysis of the distribution of network predictions in output space.

1.5.3 The hybrid Monte Carlo algorithm

The application of MCMC techniques to neural network learning as described in (Neal, 1996) is based on techniques borrowed from statistical physics for modeling physical systems. These systems consist of particles described at any point in time in terms of position, mass and momentum. A distribution over all possible states of the system is defined in which states with higher energy are less probable than those with lower energy. This is known as the canonical distribution. In many cases it is possible to sample directly from the canonical distribution using the Metropolis algorithm without going through a dynamical simulation of the physical system. However, recent work has shown that the method of dynamical simulation offers advantages as a sampling technique that sometimes make it more efficient than the direct approach (see Neal, 1993 for a review). In the case of neural network learning, it is advantageous to reformulate the problem in dynamical terms in order to speed up the rate of convergence to the posterior. The analogy between the neural network and a physical system is summarized in table 1.

Following Neal (1996, pg. 58 ff.) hybrid Monte Carlo can be described as an elaboration of "stochastic dynamics", which is an extension of Hamiltonian

Table 1 Formulation of neural network training as a dynamical simulation

SYMBOL	NEURAL NETWORK	PHYSICAL ANALOG
$\mathbf{w}$	Weight Vector	Position Vector
$\mathbf{m}$	Momentum Term Vector	Momentum Vector
t	Training Iteration	Time
ε	Step Size	Time Delta
E	Training Error	Potential Energy
K	Sum square of momentum terms	Kinetic Energy
H	$E+K$	Total Energy ($E+K$)

dynamics. In a dynamical simulation, network weights correspond to the vector defining the position of particles, and a parallel vector of momentum terms is introduced. Training error corresponds to potential energy. In order to use dynamical simulation it is necessary to be able to take the partial derivatives of this term with respect to the positions (weights). We can do this using error backpropagation (Rumelhart et al., 1986).

Hamiltonian dynamics describes changes in position and momentum that preserve the total (kinetic + potential) energy of the system. It is described by differential equations that can be simulated using the discrete update rules outlined below. The notation used is defined in table 1.

1. $m_i(t+\frac{\varepsilon}{2}) = m_i(t) - \frac{\varepsilon}{2}\frac{\delta E}{\delta w_i} w_i(t)$

2. $w_i(t+\varepsilon) = w_i(t) + \varepsilon(m_i(t+\frac{\varepsilon}{2}))$

3. $m_i(t+\varepsilon) = m_i(t+\frac{\varepsilon}{2}) - \frac{\varepsilon}{2}\frac{\delta E}{\delta w_i} w_i(t+\varepsilon)$

These are called "leapfrog" updates because first the momentum is updated by a half step, then the weights are updated by a full step, and finally the momentum is updated by another half step. The effect of the momentum terms in the neural network application is to suppress random walk behavior. This is important when sampling from distributions in which there are strong correlations, as is usually the case with neural network weights. In conventional optimization approaches to neural network learning, this has led to the use of techniques like adding a momentum term to backpropagation updates, or using conjugate gradients. In fact, simply using leapfrog updates to simulate Hamiltonian dynamics without adding the

stochastic elements described below leads to a system that is similar to backprop with momentum (Neal, 1996, pgs. 111 - 112). The momentum terms determine the kinetic energy of the system, which is calculated as half of the sum square of the terms.

This discrete approximation of Hamiltonian dynamics samples from the canonical distribution for a fixed total energy, H. However, we need to sample from the whole of the canonical distribution. The "stochastic dynamics" method (Andersen, 1980) accomplishes this by alternately sampling from the canonical distribution given a fixed total energy using Hamiltonian dynamics, and then resampling total energy by applying Gibbs sampling on the space of momentum vectors. In the Gibbs sampling phase, each of the momentum parameters is replaced in turn, according to its probability conditioned on the other momentum parameters. This can be calculated using the definitions of kinetic energy and its probability distribution:

- $K(\mathbf{m}) = \sum_{i=1}^{n} \frac{m_i^2}{2}$
- $p(K(\mathbf{m})) = \exp(-K(\mathbf{m}))$

The discrete approximation to Hamiltonian dynamics is a source of systematic error in the stochastic dynamics method. This can lead to changes to H, which would be unchanged if the simulation was exact. The error is eliminated in the hybrid Monte Carlo method (Duane, et al., 1987) through the use of the Metropolis algorithm. Still following (Neal, 1996, pg. 60 ff.), we can define this as a modified version of stochastic dynamics in which sampling proceeds through a series of "dynamical transitions". Each transition starts from the current state

$(\mathbf{w},\mathbf{m})$ and proceeds through a series of leapfrog steps after which the momentum terms are negated. This results in a new state $(\mathbf{w}^*,\mathbf{m}^*)$ that is considered as a candidate state and accepted with probability

$$\min(1,\exp(-(H(\mathbf{w}^*,\mathbf{m}^*)-H(\mathbf{w},\mathbf{m})))),$$

thus tending to reject moves to the extent that they modify the total energy. If the candidate state is rejected, the current state is unchanged. The negation of the momentum variables is required to satisfy the requirement for a symmetrical proposal distribution for the Metropolis algorithm to be valid. This procedure removes the error introduced through discrete simulation.

1.5.4 Discussion of hybrid Monte Carlo

The hybrid Monte Carlo algorithm provides a means of applying Bayesian inference to the problem of neural network learning. The next chapter will be concerned with validating this technique using simulated data. I will compare the output distributions generated by the neural network system with exact solutions for a series of simple problems. The remainder of this thesis will concern the application of this technique to a real world problem in data analysis. This will allow us to test the generalization of the system and to compare its performance with standard statistical techniques.

One concern with using any MCMC technique is the problem of diagnosing convergence. This is certainly a concern if you are working on a problem that is highly nonlinear and requires lots of training data. In this case it can even be difficult to know when to stop a conventional optimization of neural network weights. In the application discussed here, diagnosing convergence has not been a

problem. Typically, the system starts to overfit quickly (within a few minutes) and then recovers. This is because in most cases I am working with sparse and noisy data, so that the effective complexity of the model derived is not great. This means I will not in this application be testing the full power of the neural network system as a nonlinear model. This, however, has been amply demonstrated in much previous work involving neural networks, for example on the two spiral problem (Lang and Witbrock, 1988), at the Santa Fe Chaos Competitions (Weigend and Gershenfield, 1994), and at the Energy Prediction Competitions (MacKay, 1993). The focus of this work will be on the adaptive stabilization procedures inherent in the Bayesian approach, and the performance of this system given sparse data. These properties will be crucial in the context of scientific research.

A second concern with this implementation is the likelihood of the simulation becoming stuck in the neighborhood of a single local minimum of the error surface. There is no provision in the implementation to attempt to avoid this, for example through the use of simulated annealing although this was tried in an earlier implementation (Neal, 1996, pg. 65). The existing implementation is in fact likely to become "stuck". However, this is not a great concern provided that the networks used have large numbers of hidden nodes. Experience in training neural networks suggests that, although the quality of local minima can vary greatly for networks with small numbers of hidden nodes, the local minima for large networks tend to be similar to each other. The fact that the presence of many local minima in these very high dimensional optimization spaces does not in practice harm the performance of neural networks is borne out by the empirical success of the field in general, and of this implementation in particular (Neal, 1996, chapter 4).

1.5.5 Adaptive regularization

As described above, regularization in these models is achieved by setting zero mean Gaussian prior distributions on weight values. This is equivalent to the use of a sum squared "weight decay" error term in conventional neural network training. In the conventional approach, the training error would be augmented with this second term for weight error defined as:

$$C\sum_i w_i^2$$

Here C is a scaling constant that determines the amount of smoothing due to the weight decay term. Determining an appropriate value for C has generally been done heuristically or empirically through the use of a validation data set.

The equivalents of the smoothing constant in the Bayesian framework are the width parameters of the priors on weight values. These variances are given initial values, which are adapted to the data during training. In alternating cycles, first the weights are adapted to the data given the current widths of the prior distributions; then the prior widths are adapted based on the effects of training. These effects are reflected in the current weight values. If the data set is large and there are consistent patterns relating inputs to outputs, the weights are likely to be pushed far from zero towards relatively large values. Training sets that are small and noisy will allow the priors to dominate, and the weights will remain small. In the second phase of the training cycle, the weight priors are adjusted based on the effects of training. In this implementation, this is accomplished by alternating hybrid Monte Carlo updates of the weights with updates of the weight priors using Gibbs sampling.

It may seem odd that in the framework of Bayesian inference we are updating the priors based on the data! This is accomplished through a hierarchical

model definition in which the variances of the weight priors are themselves given prior distributions that can legitimately be updated after the data has been seen. It is possible to assign different priors to different groups of weights: for example the prior for the hidden to output weights should generally be different from the prior for input to hidden weights. In the following, members of one such set of weights are designated as u_i. In Neal's formulation (Neal, 1996, pg. 66 ff.), it's convenient to work with "precision" terms, τ_υ, defined as $\sigma_\upsilon^{-1/2}$, where σ_υ is the variance of the weight prior for this group. Under the assumption of zero mean independent Gaussian distributions, the prior on weights is defined as:

$$P(u_1 ... u_k \mid \tau_\upsilon) = (2\pi)^{-k/2} \tau_\upsilon^{k/2} \exp(-\tau_\upsilon \sum_i u_i^2 / 2)$$

The hyperprior over τ_υ is then given a gamma distribution with mean ω_υ, and shape parameter specified by α_υ. From this Neal derives:

$$P(\tau_\upsilon \mid u_{1...} u_k) \propto \tau_\upsilon^{(\alpha_\upsilon + k)/2 - 1} \exp(-\tau_\upsilon (\alpha_\upsilon / \omega_\upsilon + \sum_i u_i^2) / 2) .$$

This PDF for the weight prior precisions conditioned on the current weight values provides the basis for Gibbs sampling updates of the weight priors in which each precision is replaced in turn based on its PDF with the other precisions fixed.

1.5.6 Bayesian model selection

A key problem in applying neural networks is determining network architecture. Even if we accept the usual choice of a single hidden layer fully connected to the inputs and outputs, we have to decide on the number of hidden units. This is a special case of determining model complexity. Conventional wisdom has long been that model complexity must be chosen based partly on the amount and quality of the available training data. Neural networks with many hidden nodes, like any model containing large numbers of adjustable parameters, were considered inappropriate choices if the data set was small; overfitting was considered to be inevitable. This belief was reinforced by David MacKay's work on the application of Bayesian inference to neural network training through the use of analytical approximations. MacKay (1992a) advanced arguments for an "Occam's razor" principle favoring simpler models in the model selection process. He also found empirically that model performance declined when the number of hidden nodes grew too large (MacKay, 1992b). This is a familiar experience in the application of neural networks.

Radford Neal's work represents a radical departure in this respect. He has pointed out that there is no basis in Bayesian theory for this approach to model selection:

> "From a Bayesian perspective, adjusting the complexity of the model based on the amount of training data makes no sense. A Bayesian defines a model, selects a prior; and then makes predictions. There is no provision in the Bayesian framework for changing the model or the prior depending on how much data was collected. If the model and prior are correct for a thousand observations, they are correct for ten observations as well (though the impact of using an incorrect prior might be more serious with fewer observations)." (Neal, 1996)

An important feature of Neal's neural network system in this respect is an automatic scaling of the prior distributions for the weights out of the hidden nodes based on the number of hidden nodes. He has demonstrated (Neal, 1996, pg. 32) that the variance of neural network output values scales as $H\sigma_u^2$, where H is the number of hidden units, and σ_u is the standard deviation of the prior for the weights out of the hidden units. Therefore by scaling σ_v as $H^{-1/2}$, the variance of the outputs of the function implied by the prior remains constant for any number of hidden nodes. Neal has reported results for networks with 6, 8, 16 and 32 hidden nodes on the same problem (Robot Arm) for which MacKay had reported overfitting by large networks. In these tests there is no consistent pattern of overfitting by the larger networks, and there is a clear pattern of underfitting by the smallest. The contrast with MacKay's results may suggest that the overfitting that he reported was due to a breakdown in the Gaussian approximation given networks with large numbers of parameters.

In this application I have found that Neal's scheme for scaling the prior based on the number of hidden nodes works well. In fact, after some initial experimentation I completely stopped worrying about the numbers of hidden nodes beyond making sure that there were "plenty". For the two class problems described later in this thesis I have used 8 hidden nodes. For the three class problems I have used 12 hidden nodes. The extensive model validation reported in the next two chapters has ensured that this has not led to overfitting. The significance of this contribution of Neal's work should not be underrated. In the past I have either used a validation set or cross validation to determine the number of hidden nodes. This is a time consuming process, and even when cross validation is employed, some of the information in the data is "used up" in determining the model. Particularly in a

scientific study, this will raise doubts about the validity of results obtained using the model. This was an important factor in my decision to use Bayesian neural networks in this study.

1.5.7 The application of Bayesian neural networks in a scientific context

One area where the merits of Bayesian neural networks have not been demonstrated until now is scientific research. Even Radford Neal (1996, pg. 7) has questioned their use in this context despite their undoubted effectiveness in "the messy contexts typical of engineering applications". My decision to apply these models to the analysis of physiological time series data contained in the Edinburgh head-injury database was in part motivated by developments in head injury research. Clinical researchers have developed numerous hypotheses regarding the physiological sequelae of traumatic brain injury; e.g. see (Rosner, 1985). These have primarily been based on single case studies and short series of patients. It had been hoped that large scale data acquisition projects like the one in Edinburgh would lend support to some of these, but results to this point had been disappointing. There are several possible explanations for this including, of course, deficiencies in the hypotheses themselves or of the data collection process: missing data, and failure to control for confounding factors among other problems. Another possible explanation that has been raised, however, is that the statistical analyses being employed may be too simplistic and therefore miss significant features of the data. This seemed an ideal application to test the claims being made for the Bayesian framework for neural network modeling.

The application of Bayesian neural networks in this new context has required several technical innovations. As argued in chapter two, a key problem in this application is the representation of probability densities in output space based on the discrete Monte Carlo approximations. I have developed techniques for two and three class problems based on kernel density estimation that accomplish this. I have also when possible employed numerical techniques for the exact derivation of these probability densities. This has enabled comparison of the neural network estimates with the actual densities. This in turn allowed a detailed examination of the effects of various choices for the prior distribution on the performance of the model for a range of training set sizes. It is demonstrated that approximately uniform priors on output functions are desirable for this task, and a procedure for finding such priors is explained. The density estimation techniques employed also led to a natural way of assigning confidence regions around predictions. Input standardization has also proved to be important. It was necessary to center the range of interest for each input precisely under the sigmoid transformation, and a new procedure is demonstrated that avoids problems introduced by using the usual method of subtracting the mean and dividing by the standard deviation.

1.6 Alternative technical approaches

1.6.1 Gaussian Processes

The previous section on Bayesian neural networks describes a scaling for the priors of hidden to output weights of $H^{-1/2}$ with H being the number of hidden units (Neal, 1996). This result led Radford Neal to investigate the nature of functions implied in the limit of networks with infinite numbers of hidden units assuming this scaling and

independent Gaussian priors on network parameters. He found that these functions have the property that the joint distribution of output values produced for any finite set of input vectors is multivariate Gaussian. This means they belong to a class of functions known as Gaussian processes. This result has led to a renewal of interest in Gaussian processes as techniques for regression and classification (see MacKay, 1998 for a review). The appeal of the Gaussian process approach to modeling is that it dispenses with the weight parameters used by neural networks and directly models a space of functions relating inputs and outputs. I briefly tried some experiments with Gaussian processes, which are available as an option using Radford Neal's software. I did not get good results, most likely because of my lack of experience with this method. My decision not to pursue this line of research was purely pragmatic. By this time I had already started getting promising results with Bayesian neural networks. Since my project was a very applied one I decided to go with what was working, rather than a theoretically appealing alternative that was likely to produce similar results in practice. I do, however, think that Gaussian processes will be an interesting area for future research.

1.6.2 Support Vector Machines

Support vector machines (SVM) (Cortes and Vapnik, 1995) provide an alternative to neural networks for the problem of nonlinear classification. Given a training set for a two class problem, the "maximal margin hyperplane" is determined. This is defined as the linear surface that separates the classes with a maximal distance from the "support vectors": those class instances that lie closest to the decision surface. This approach can be generalized to

nonlinear problems by projecting the data into a higher dimensional feature space using nonlinear kernel functions in such a way as to make them linearly separable in that space. One principle of the SVM framework is to produce classifiers with minimal VC dimension, which will generalize well to new data (Vapnik, 1998).

In principle, the SVM approach satisfies the criteria I have set for this application. It is a nonlinear classification technique that adapts based on the amount and quality of available training data. SVM have already been used with great success in many applications. As with Gaussian processes, my decision not to try this approach was partly pragmatic. Faced with a very applied project, I selected a more familiar method. I would say that the interpretation of model outputs is more straightforward using Bayesian neural networks. On the other hand interpretation in terms of model complexity is probably more straightforward in the SVM framework. Also, recent work has suggested an equivalence between SVM and the Bayesian approach (Cristianini and Shawe-Taylor, 1999). I think that the application of SVM to this intensive care data set would be an interesting area for future research.

1.6.3 Neural networks with bootstrap

As discussed above, the nature of this project didn't lend itself to conventional neural network approaches because of sparse data and the need for a clear mathematical interpretation of results. Besides the Bayesian framework, one other method discussed in the literature seemed promising. This was work by Baxt and White (1995) that assessed the prognostic value of clinical features using a neural network and applied a bootstrap analysis to assign confidence intervals to their

estimates regarding the relative importance of these variables as prognostic indicators. The bootstrap varies a database by randomly sampling training cases from it with replacement. This has the effect that for any given training run, some training cases are represented more than once, and some are not selected at all. This permits an analysis of how robust and generalizable results are given the amount of available training data. It's a sort of analog to MCMC sampling on the space of possible models that operates by varying the database rather than studying model variations directly. I prefer the Bayesian approach because the methodology and the interpretation of results are more straightforward. The recent advances in the application of Bayesian inference to complex models which have been discussed above, combined with the availability of ever increasing computer power have made a full Bayesian analysis much more accessible than it was a few years ago. This has led even some of those who have applied the bootstrap very effectively to refer to it as a "poor man's Bayes" (L. Breiman, personal communication).

1.6.4 Optimization on choice of input features

One general approach to analyzing a complex database is to focus on the problem of feature extraction. Previous work with head injury data has suggested that feature selection may be more important than choice of modeling technique (Titterington et al., 1981). By settling on a modeling technique that is not very computationally demanding, the task reduces to an optimization problem in the choice of features. This could be approached by applying standard techniques of variable selection (forward selection or backwards elimination), or through more advanced techniques such as simulated annealing and genetic algorithms. My reason for rejecting this

approach was primarily that it had, to a very large extent, already had been done. Variable selection had early on been used to evaluate simple prognostic features following head injury (Braakman, et al., 1980). Most notably (from my standpoint), it has been applied to a set of 187 candidate features automatically generated from physiological data from head-injured patients (Marmarou et al., 1991). This study evaluated various methods of feature extraction, and the clinical significance of the features. Variable selection has also been used to evaluate a more limited set of candidate physiological features extracted from the same Edinburgh head injury database which will be the subject of this thesis (Signorini et al. 1999b). These studies will be reviewed in detail in chapter four.

1.6.5 Decision Trees and Bagging

The automatic induction of decision trees explores the space of possible input features and produces a classifier at the same time. Trees are produced by recursively splitting a data set according to a series of binary features. Features are generated from the variables available as inputs to the classifier. They are evaluated according to their predictive value in relation to a set of target variables. In this application the inputs are demographic and clinical indicators available for a set of patients, and the targets are the eventual outcome for those patients: for example survival or death. An example of such a feature might be whether or not the age of the patient is over 40. Features are selected one at a time on information theoretic grounds. The selection process varies, but a typical criterion would be that using the feature to divide the training set into two classes maximizes within class entropy and minimizes between class entropy. Once a feature is selected, the process continues

by dividing the subgroups in the same manner. Typically the tree is grown until the leaf nodes all contain one single case. Then it is pruned back to produce a simpler tree with more general applicability. The pruning can be done automatically or with human help. Intervention in this way by a human expert can permit exploration of the data and allow the introduction of expert knowledge. It is also possible to construct a partial tree manually and complete it using the data driven methods.

The great advantage of this approach is that the output of the classifier has an immediate clinical interpretation. Inspection of decision trees may confirm or challenge expert opinion. A problem with decision trees is that given complex real world problems, they do not often produce classifiers that are as reliable as the best statistical systems, or neural networks. This is because of the way they combine evidence. The algorithm described here is a "greedy" algorithm. That is it selects features one at a time, choosing the locally optimal feature at each splitting point. This is not likely to produce the globally optimal tree, even if a tree structured classifier is appropriate. This reliance on a somewhat arbitrary series of decisions is a particular problem when dealing with sparse and noisy data, as will be the case in this application.

There is a way of improving the performance of decision tree techniques to the point that they *are* competitive with the best classification systems. This is through a technique called "bagging" for "bootstrap aggregation" (Breiman, 1996). This involves generating a large number of trees and averaging over their predictions. The training set for each tree is varied through bootstrap resampling. That is, N cases are produced for each tree by selecting from the available training data *with replacement.* This means that in each training set some cases will be

selected more than once, and some won't be selected at all. By varying the data sets in this way you get a variety of trees, and by averaging over their predictions you get a classifier with much better generalization. However, the advantage of the decision tree approach now becomes a liability. Since the data is being processed by a multitude of trees, the classifier is *only* useful as a black box source of predictions. It is not even straightforward to investigate the relationship between inputs and outputs of the averaged model, since each tree utilizes its own choice of input features.

1.6.6 General rule discovery and boosting

Decision trees are a special case of systems that discover rules in data. More unconstrained systems utilize general graphical structures, or even construct programs that can be used for classification (Holland, 1986). Often these are based on heuristic principles or are designed by analogy to economic or ecological systems, which makes it difficult to understand their operation or justify their decisions to domain experts. It might seem that explanation should be easy given systems that discover rules, but this is usually not the case. In fact the problem of "credit assignment", i.e. determining which set of rules was responsible for the success or failure of the system in any given instance, can be very difficult, and has been the subject of much research. Still, some of these techniques have been shown to be effective in problems involving optimization in high dimensional feature spaces (e.g. see Feng and Michie, 1994, Sedbrook et al., 1991). The use of these systems will become more attractive in medical applications as larger databases become available.

Boosting (Schapire, 1999) is a relatively new technique for rule combination that is being used very effectively in a variety of applications. Boosting starts with the development of a "weak learning" algorithm. This is an automatic procedure for generating "rules of thumb" given a training set. These rules of thumb are simple ways of classifying instances. For example: "the patient will die if blood pressure is less than 80 more than 30% of the time". These can be "weak" rules in that their performance as classifiers (individually) need be only slightly better than random guessing. As these rules are generated they are combined with previously accepted rules using a numerical weighting scheme. The training set is modified as rule selection proceeds by dropping out cases that are confidently and correctly classified by existing rules. Again, I suspect that in the application described in this thesis, small sample size would be a problem, as would lack of explanatory power due to the credit assignment problem. However, Boosting might be an interesting area for future research.

1.7 Statistical prognostic modeling and head injury research

Head injury research has been an active area for many years for research involving large scale data collection and analysis from a number of different perspectives. Early work focused on quantifying basic clinical information such as depth of coma and the quality of patient outcome. This work led to the definition of the Glasgow Coma Scale, (Teasdale, G., Jennett, B., 1974), and the Glasgow Outcome Scale, (Jennett and Bond, 1975): clinical tools that have become world-wide standards. These are defined in tables 2 and 3. This led to the development of the "three

country database" that combined clinical indicators of injury severity and quality of recovery with demographic data for 1000 head- injured patients in Britain, the Netherlands, and the Unites States (Jennett, et al., 1977). This database was the subject of a major study in applied statistics in which several different modelling techniques were compared (Titterington et al., 1981). This led to the development of the Glasgow prediction program (Barlow et al., 1984), which implemented a naïve Bayes model for predicting patient outcome following head injury. The work in Glasgow continued with a series of studies comparing the performance of the

Table 2 The Glasgow Coma Scale (GCS) is used world-wide to quantify depth of coma. It is calculated as the sum of the three component scores, giving a minimum score of 3, and a maximum of 15.

GLASGOW COMA SCALE		
Eye Opening	**Motor Response**	**Verbal Response**
1 – None	1 - None	1 – none
2 – Responsive to pain	2 - Extension	2 - Sounds only
3 – Responsive to command	3 - Abnormal Flexion	3 - Words only
4 – Spontaneous	4 - Normal Flexion	4 - Confused speech
	5 - Localises Pain	5 - Orientated speech
	6 - Obeys Commands	

Table 3 The Glasgow Outcome Scale is a scale from 1 to 6 quantifying the quality of patient outcome following brain injury. This is the standard used in head-injury research.

GLASGOW OUTCOME SCALE	
Score	**Definition**
1 – Death	
2 – Persistent Vegetative State (PVS)	Patient remains in coma
3 – Severe Disability	Requiring help with at least one daily activity
4 – Moderate Disability	Self-caring, but not back to previous level of function
5 – Good Outcome	Back to previous level of function

prediction program with that of clinical practitioners in predicting outcome for patients, and exploring the uses of this kind of program in clinical practice (Teasdale, 1981).

In the United States, a joint head injury study was launched which produced the Traumatic Coma Data Bank (TCDB) (Foulkes., et al., 1991) . This database included end hour recordings of physiological parameters such as arterial blood pressure and intracranial pressure transcribed from the nurses chart. This enabled researchers to study possible models of outcome following head injury that incorporated data from the Intensive Care Unit (ICU) itself, as well as severity of injury and outcome information (Marmarou et al, 1991).

1.8 Applications of artificial intelligence in intensive care

Intensive care would appear to be an ideal field for the application of computer technology in general and artificial intelligence in particular. The patient bedspace in a modern ICU is surrounded by electronic monitoring devices displaying measurements being updated, sometimes several times per second. In most units this information is periodically recorded and collated manually onto a paper nursing chart: a laborious and error prone process. It has been demonstrated that much significant information regarding patient physiology is lost because this can only realistically be done on an hourly or half hourly basis (Corrie, et al., 1993). Recent years have seen this situation begin to change with advances in data communication standards and software systems for data acquisition in the intensive care environment. Some intensive care units can now claim to be truly "paper-free". As systems become more generally available and standardised, it will be possible to

begin to address questions and problems that have persisted for decades regarding the management of patients in intensive care. Some of these areas will be discussed below.

1.8.1 Clinical audit of patient management

A fundamental application of computer technology in intensive care will be quantifying differences in patient management between different centres, and the effects of these differences in patient care. It is generally acknowledged that these differences are significant, but it is not currently possible to say exactly what the differences are. Detailed, standardised data collection recording patient physiology, drug administrations and surgical interventions will enable quantitative studies and ultimately help resolve long standing debates regarding the effects of these different policies.

1.8.2 Intelligent multiparameter alarms

An obvious problem to anyone visiting an intensive care unit is the frequency with which audible alarms are generated by the patient monitors. Each ICU has its own policy regarding the thresholds used for each monitor, and how they are to be handled. Nevertheless, when a unit is very busy it is sometimes evident to the most casual observer that it is completely impossible for the staff to attend to all of the alarms resonating in the background as they try to focus on the most critical problems of the moment. The need for automatic systems that are based on the overall state of the patient rather than thresholds on single parameters, and which weigh some alarms more heavily than others is very great. This has been recognised

for a long time (for a review see Fackler, 1998) but the absence of networked computerised monitoring in most ICU's has impeded progress in this area. From an Artificial Intelligence standpoint, this is a problem in data fusion, and it could be an important area for research and applications.

1.8.3 Early warning systems

As databases of physiological time series data become available for study and statistical analysis it will be possible to develop a new generation of predictive models for use online in the ICU. These models could be used to generate warnings regarding the onset of adverse events to allow early intervention. For example, it has been shown that features extracted from EEG recording can be used to predict vasospasm following subarachnoid haemorrhage (Vespa et al.1997). In some cases, it has been demonstrated that there is *no* obvious way to detect certain adverse clinical events even given the most intensive patient monitoring in intensive care (Andrews et al., 1996). This may indicate that there is scope for sophisticated pattern recognition techniques in this area.

1.8.4 Decision Support

The staff in the neuro ICU must be prepared to respond to crisis situations at any time, day or night. It is important, both for patient safety and for the morale of the staff, that all of these situations be handled efficiently and in a consistent manner. It is not possible to maintain rigid guidelines in such a complex environment, but when staff diverge from general guidelines, they should be aware that they are doing so, and be able to justify their actions.

Progressively more detailed patient histories are being entered into computers in real time, recording patient treatment and response to therapy. This has enabled the beginning of work on automatic systems to support clinical decision making in the ICU (Ambroso, et al., 1992). This in turn has led to a recognition of the importance of the representation of expert knowledge and its efficient retrieval: familiar territory for the researcher in artificial intelligence. There has been much work on standardising medical knowledge representation systems, but this has been hindered by the wide range of specialist environments and the rapidly changing nature of these fields (Coiera, 1995). This would be an ideal application area for work in expert systems and knowledge representation.

1.8.5 Pattern recognition and machine learning

The application areas discussed so far deal with knowledge representation and management. Possibly the most important area for current research involving computers and intensive care is extracting new understanding from the masses of data being accumulated by automatic systems around the world. Interest in automatic systems that assist in analysing these kinds of complex databases has grown with the information explosion, and has become an important subspecialty of artificial intelligence and statistics (Michie et al., 1994). One common approach to this problem is to apply techniques for the automatic induction of decision trees. This has the advantage of sometimes producing results that have a direct clinical interpretation. A second approach is to develop complex nonlinear models, e.g. neural networks, and analyse the behaviour of the models (e.g. Baxt and White, 1995). This is the approach adopted in this thesis.

1.9 Summary

The information explosion in intensive care makes it a rich field for applications of Artificial Intelligence. Currently one of the most pressing needs is for the development of machine learning and pattern recognition techniques to help understand the significance of the large and complex data sets that are rapidly being accumulated around the world. The study of patient management following head injury is well advanced in this respect. Several large projects have applied statistical techniques to the analysis of databases consisting of demographic data, clinical indicators, and outcome scores. Now databases are being collected that incorporate in addition detailed physiological time series data collected during intensive care.

The Bayesian framework for neural network learning leads to systems that generalise well beyond their training set. The network outputs have a clear interpretation. Many problems with specifying and validating these complex nonlinear models have natural solutions in this framework. These advances make them an attractive alternative for the analysis of the complex data sets now being collected by medical researchers. This thesis describes the use of Bayesian neural networks to model data contained in one such database. It will present the results of these experiments and their implications for the management of patients who have suffered severe head trauma. This will require several technical innovations in the implementation of Bayesian neural networks, and demonstrate that they have a role to play in medical research.

Chapter 2

Modeling Two Kinds of Uncertainty

This chapter will describe the characteristics desired in a modeling technique to apply to the analysis of this complex medical data set. It will be shown that Bayesian neural networks possess these. The problem of specifying a prior distribution on network parameters will be addressed. Then Bayesian neural networks will be applied to a series of simple problems using simulated data, and it will be shown that they correctly model the probability distributions of the targets conditioned on the training set. These distributions scale appropriately with the size and consistency of the data set. A few simple problems requiring generalization are presented, although this problem will be largely deferred until the next chapter. A new procedure for normalizing input data is demonstrated that avoids problems introduced by the standard procedure in the presence of sparse data.

2.1 A dilemma in data analysis

Medical data sets vary along two mutually constraining dimensions that define the quality of the information they contain. Some medical studies are based on a few simple clinical and demographic indicators that are relatively easy to collect. These can include large numbers of patients without over-stretching their resources. Other studies go into more depth, collecting large amounts of data per patient. These may require expensive monitoring devices and special care on the part of the investigators to ensure that the data are collected and validated. Inevitably these studies are on a smaller scale in terms of numbers of patients. Often a study will produce a mixed

database in which there are varying amounts of information per patient. This poses a dilemma for the data analyst. If you select a data set with a few simple predictors and a large number of patients, you may be discarding your most significant data. On the other hand, as you include more data per patient, your patient numbers get smaller. Ultimately this dilemma seems artificial. Surely in these cases the aim should be to develop a variety of models, each appropriately characterized in terms of uncertainty due to the amount of available information per patient, and also the uncertainty due to patient numbers. This would allow medical researchers to evaluate their current database and better focus future data collection efforts.

Given large patient numbers, estimating uncertainty attributable to information not captured by the input variables to a neural network is relatively straightforward. It is well known that the use of the softmax activation function and cross-entropy error term leads to networks that model conditional probabilities (Bishop, 1995 pg. 230 ff.). This provides a measure of uncertainty due to limitations on the amount of information for each patient. Measuring uncertainty due to small patient numbers is more problematic. Baxt and White (1995) proposed an approach based on results from sampling theory. Here I present an alternative approach based on Bayesian inference. This framework has the advantage of producing systems of networks whose outputs can be directly interpreted as the desired probability density functions

.

2.2 Bayesian inference and neural networks

Neural Networks are usually optimized to fit a training set, producing the familiar error-reduction curve that asymptotically converges to a local minimum of the error. This process can be described in the framework of Bayesian inference. Before the data arrived our prior beliefs lead us to select the form of the model and set the initial parameter values. Then the parameter values are modified in the light of the data, giving rise to the optimized network. There is a glaring deficiency in this procedure. No amount of data can tell us with a probability of one that our optimized neural network is the true model to the exclusion of all others. Rather, the data support to a greater or lesser extent an infinite number of possible models. It is especially important to recognize this principle when dealing with a nonlinear model with large numbers of adjustable parameters like a neural network. Rather than selecting the single model best fitting our particular data set, the Bayesian approach is to approximate the posterior distribution on the space of possible models given the data. This framework produces models that generalize better to new data, and also allows us to estimate model uncertainty due to limitations of the available data. If the data set is large and consistent, then the posterior distribution on network outputs will be tightly constrained. Small noisy data sets will give rise to broad posterior distributions representing a high degree of uncertainty. David MacKay has given a detailed exposition of the application of Bayesian inference to the development of neural network models (MacKay, 1992a - c).

One approach to approximating this probability distribution is to generate a series of networks in such a way that any particular network is produced with the

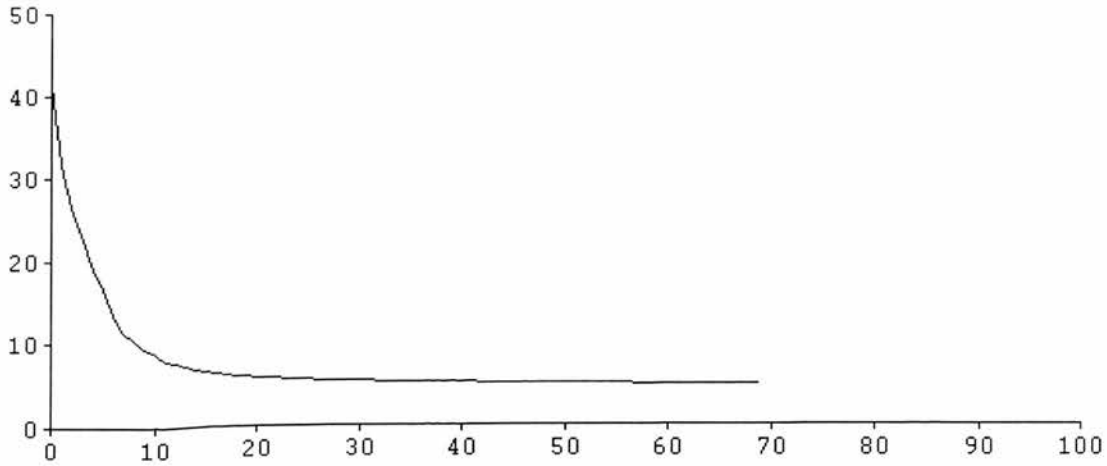


Figure 1 Error trace for a neural network being trained using a conventional gradient descent algorithm. The vertical axis is error, and the horizontal training iteration

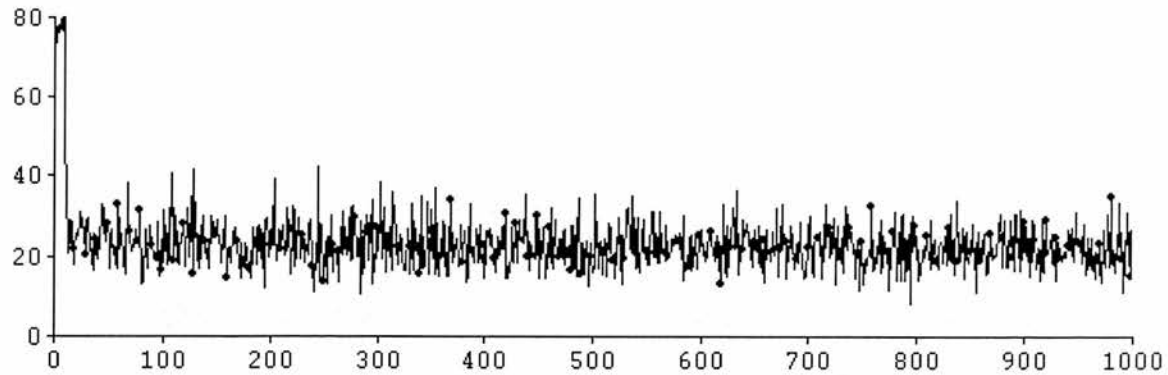


Figure 2 Error trace of neural network training using Radford Neal's Monte Carlo simulator. Gradient information is used in conjunction with a stochastic process. The vertical axis again represents the error term, although this is not directly comparable to Figure 1, because the network was trained on a different problem.

correct posterior probability. This amounts to a discrete approximation to the posterior density function. This approach has been used in neural networks, notably by Radford Neal (Neal, 1996). Neal uses techniques adapted from statistical physics to generate Monte Carlo Markov chain (MCMC) simulations which are guaranteed to converge to the correct equilibrium distribution. Figure 1 is a trace of a typical error reduction curve for conventional neural network training, while figure 2 is a trace of an MCMC simulation. This trace and the simulations presented in the remainder of

this thesis were generated by a system based on Radford Neal's software. Unlike conventional neural network training, the Monte Carlo system does not proceed smoothly through error space towards a single parameter set, but rather explores the posterior distribution in model space indefinitely, producing a progressively more accurate approximation to the true posterior. Each dot in the graph in figure 2 represents a neural network saved out for use in making predictions. The output of this system is then the distribution of outputs from these networks. We will see detailed examples of how this works in practice later in this chapter.

2.3 Interpreting the outputs of individual networks

When a neural network is trained using the "softmax", or multilogistic, activation function and cross-entropy error term, the outputs can be interpreted as conditional probabilities in a classification task (Bishop, 1995). In the limit of an infinite amount of data, the network outputs converge to the probabilities of the targets conditioned on the input vector. These probabilities represent the uncertainty in our predictions due to information about the targets not captured by the inputs. As we collect more data, this uncertainty will not go away; that is, the probabilities will not go to one and zero. However, our estimates of the conditional probabilities will get progressively better.

To illustrate this idea, I trained a conventional neural network using the softmax activation function and cross entropy error on the legendary XOR problem. I added random noise by flipping 10% of the targets to the opposite category for each

Table 1 Approximations generated using the softmax error term

Inputs	Prob. Target = 1	Estimated Prob.
0 0	0.1	0.098
0 1	0.9	0.902
1 0	0.9	0.902
1 1	0.1	0.098

of the four possible input vector types. The training set consisted of 120 examples. The results are summarized in table 1. Thus, given a large amount of training data, the conventional network converges to the correct probabilities for the outputs conditioned on the input vectors. This quantifies the uncertainty in our predictions due to information not captured by the input vector or due to random noise. It does not, however, provide any information regarding uncertainty due to the amount of available training data.

In this case and throughout I've used a network with 8 hidden units and two output units. Of course, this is extravagant for XOR, but when I deal with real data, I'll be interested in problems with more than two output categories, and it will be necessary to specify a generous number of hidden units to ensure that the network is capable of modeling the structure of the data. I've chosen to use one-of-C output coding and a large number of hidden units to better simulate that situation.

2.4 Interpreting the MCMC output distribution

In the previous section we saw that neural network outputs can be interpreted as conditional probabilities given their inputs. In terms of a medical database, this provides a measure of the predictive power of the data selected for each patient. For

example, we might train a network on selected data values such as age, admission blood pressure, and so on. The output categories might be assessments of outcome: good outcome, poor outcome, and death. If the trained network typically produces outputs that are close to the arrival rates of the three output categories, we would interpret this to mean that the input variables selected were not predictive of outcome. On the other hand, if the output probabilities are typically close to 1 and 0, we would believe that the input variables allowed us to predict outcome with considerable certainty. The quality of these probability estimates, however, depends crucially on the size of the available training set. It is this measure of uncertainty that we can infer from the properties of the probability *distribution* estimated in the framework of Bayesian inference using Monte Carlo Markov chain methods. In the following sections I'll examine the performance of the MCMC approach in a variety of simple contexts. This will serve to illustrate the estimation of uncertainty due to sample size, and to verify the accuracy of this implementation.

In the following figures the distribution over network outputs is shown as a rug on the x-axis. Each tic-mark represents the probability estimate of one particular network. Above this line is a continuous plot which is a Parzen window density estimate of the output distribution. See appendix A.1 for details of the density estimator.

2.5 Calibrating the prior distribution

In previous work on applying Bayesian inference to neural network development, the prior distribution on network weights has typically been used to express a preference for networks with smaller weights. This is essentially a form of function

regularization closely related to the use of weight decay (MacKay, 1992c). Figure 3 shows a sample of outputs from 200 networks drawn from a typical weight-decay prior. The networks have two inputs that have both been set to one. This is the output of the system in the absence of any training data. Under the interpretation I am advocating here, this is wrong. The system is telling us that before it has seen the data it has a fairly strong preference for uncertain networks; it is already fairly certain that it is uncertain! I would prefer that uncertainty be expressed by spreading predictions evenly on the range [0 1]. A uniform prior on network outputs would express the belief that in the absence of any data, the predictive power of the input variables is simply unknown.

At first I thought that this could be accomplished simply by specifying a very broad prior. Unfortunately, the problem is not this easy. Figure 4 shows the result of increasing the standard deviation of the previous prior by a factor of 100. The original preference for uncertain networks has been replaced with a preference for networks that make highly certain predictions, which is even worse. The reason for this is that a large proportion of networks included in the high probability density region of a very broad prior will be dominated by very large weights. This leads to the prevalence of networks that predict with great certainty. As has been pointed out elsewhere (Wolpert 1994, Neal 1996 chapter 2), a distribution over network parameters is not the same as a distribution over network outputs.

There are probably better ways of designing an approximately uniform prior over outputs for classification tasks, but for now I have simply proceeded by

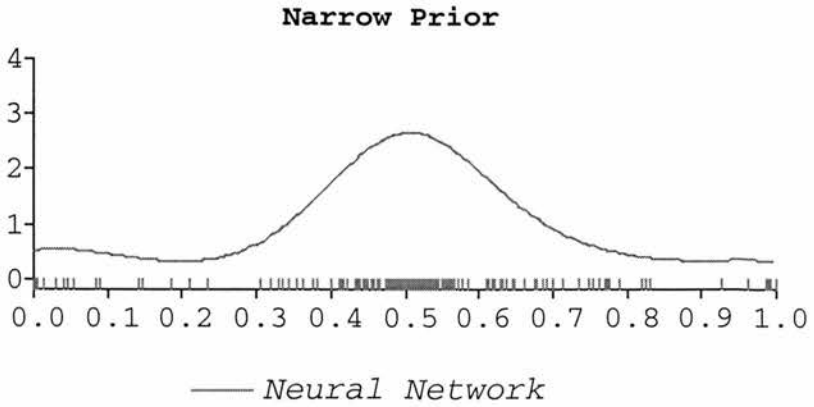


Figure 3 The outputs of 200 networks generated from a typical weight decay prior. The tic marks on the lower line are the predictions made by the 200 networks. The continuous distribution is calculated from the discrete distribution using kernel density estimation. In this and the two following figures the networks have two inputs both of which were set to one.

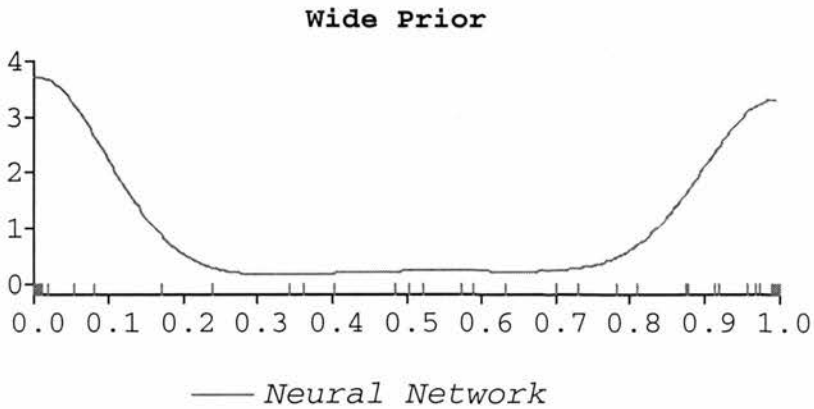


Figure 4 The result of multiplying the standard deviation of the above prior by 100

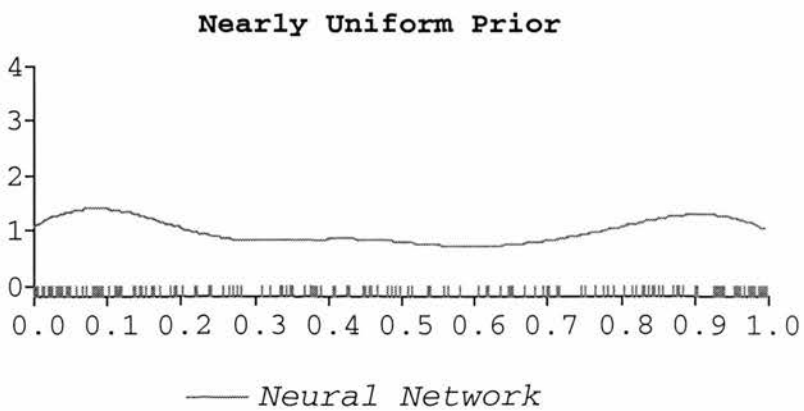


Figure 5 An approximately uniform prior on network outputs

experimenting with various prior widths until I found one that distributes its predictions fairly evenly. Since this is a means of expressing reasonable prior beliefs without reference to the data, this should not be regarded as parameter tweaking. Figure 5 represents a sample from the prior used in the following simulations. Because of the effect of the logistic function on the network outputs, it is difficult to completely avoid some bunching at the extremes of the distribution. However, since this is such a weak prior, one would hope that even a small amount of data will overcome its deficiencies.

2.6 Modeling Bernoulli trials

The previous section described the behavior of the system in the absence of any data. Now, taking a cautious step forward, we will look at how the system behaves when the inputs have fixed values, and are therefore irrelevant. The only data of interest is the value of the target variable, which is binary, and takes the value 1 with some fixed probability P. It is as if we had done some number of tests, and the only data we had collected was the outcome of each test. Based on this data, we want to estimate the probability of a positive outcome on future tests. For any given estimate of this probability we can calculate the likelihood of our data. If we have performed N tests and observed M positive results, the probability of this result for a given value of P *is:*

$$\binom{N}{M} P^M (1-P)^{N-M}$$

We can form a numerical approximation of the probability density function for P conditioned on the observed number of positive results and the number of trials by

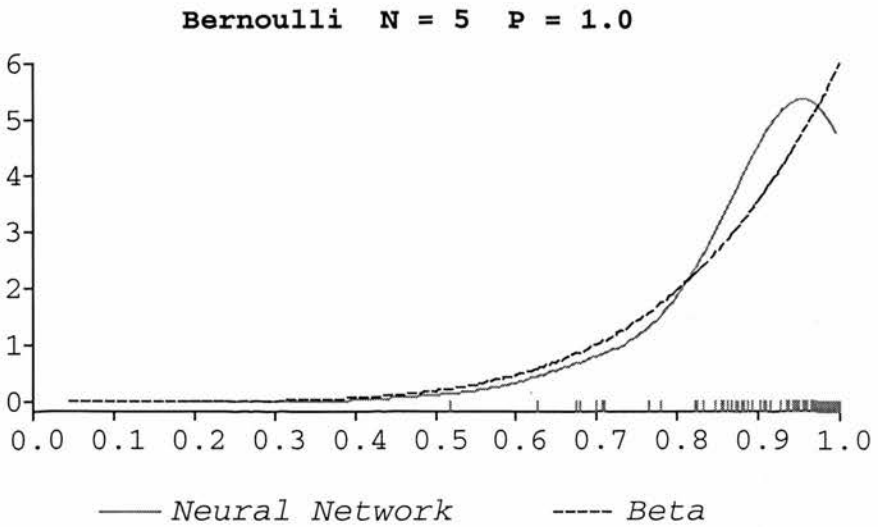


Figure 6 "Bernoulli trial" inputs and outputs: N = 5, P = 1.0

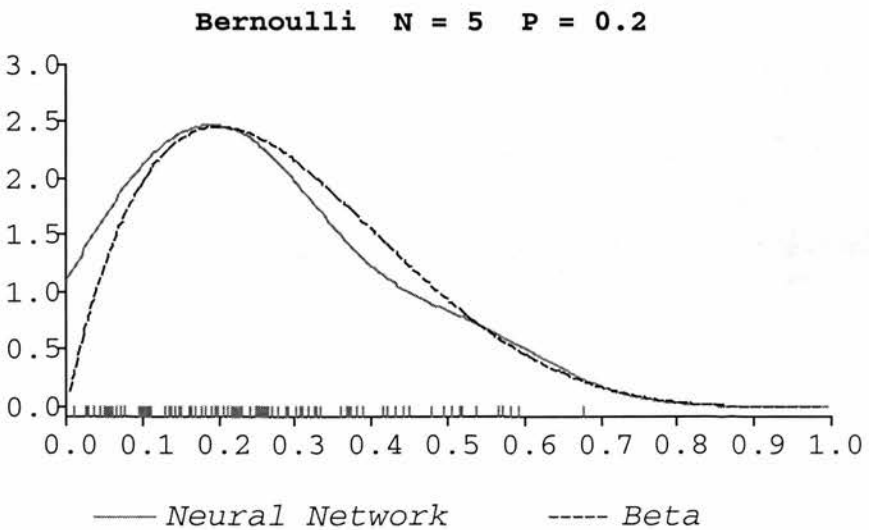


Figure 7 "Bernoulli trial" inputs and outputs: N = 5, P = 0.2

computing the likelihood of the data on a grid on the range [0 1] and rescaling so that the area under the curve equals one. This is a numerical approach to computing a beta function. In figures 6 through 9 I use this as a means of generating reference distributions to see how closely the MCMC output distribution approximates the true probability density functions. The exact distributions (beta functions) are shown as

dotted lines, while the Parzen window estimates of the MCMC output densities are shown as solid lines. The settings we have used in defining the MCMC simulations are defined in appendix A.3.

I have found that this system provides good approximations over a wide range of sample sizes and probabilities. Distortions from the true distributions are largely predictable from the distortion of the prior from a uniform distribution. The prior is slightly bunched at the extremes, and here the predictions are slightly pushed towards the extremes. A near perfect fit might be expected if a truly uniform prior could be devised. Nevertheless, these approximations are entirely adequate for this application. Figures 8, and 9 are examples from the larger sample sizes showing how the confidence tightens up as more data is collected.

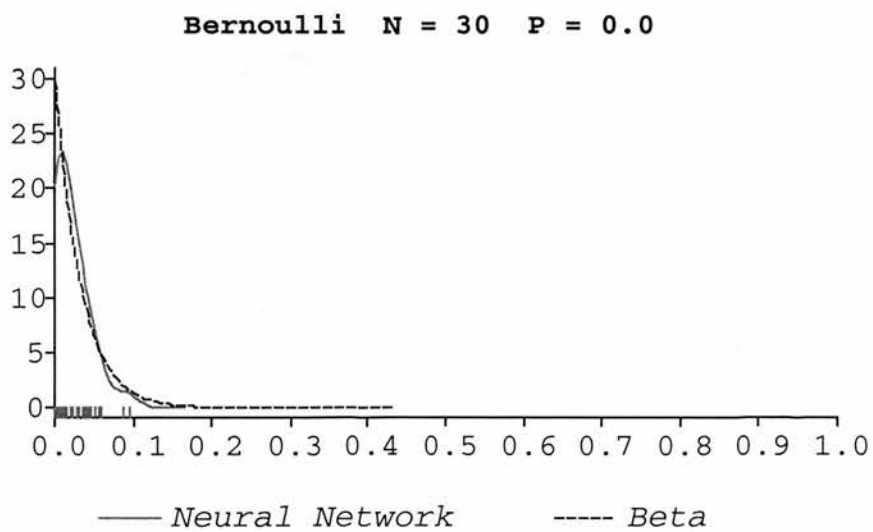


Figure 8"Bernoulli trial" inputs and outputs: N = 30, P = 0

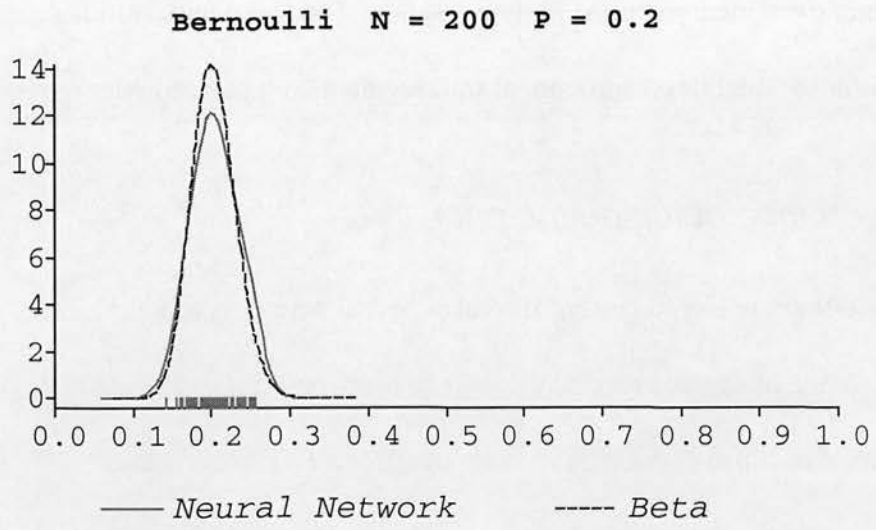


Figure 9 "Bernoulli trial" inputs and outputs: N = 200, P = 0.3

2.7 Interpreting the output distributions

These results demonstrate that the outputs generated by these MCMC simulations can be interpreted as approximations to probability density functions that capture our two kinds of uncertainty. The horizontal axis represents uncertainty due to the (lack of) predictive power of the input variables. Values located near the center (0.5) represent high uncertainty, and values near the extremes (0 or 1) represent near certainty. The shape of this distribution represents uncertainty regarding that estimate due to sample size. This simple example has allowed us to compare the estimates produced by the neural network to the true PDF's. Of course, one would never use such a complex, computer-intensive modeling technique to solve a problem like this. It is reassuring, however, that the system is able to capture the very simple structure of this problem across a broad range of sample sizes. The tendency of neural

networks to slip into overfitted solutions in the presence of sparse data has in the past limited their usefulness when large amounts of training data are not available.

2.8 The importance of tuning the prior

In section 2.5 I said that the key to getting Bayesian neural networks to work properly in the presence of sparse and noisy data is to tune the prior distribution so that it approximates a uniform distribution in output space. To illustrate this principle I have repeated the "Bernoulli trial" experiments of section 2.5 using first the carefully tuned "uniform" prior and then a more "standard" prior that favors uncertain network predictions. For the uniform prior I have used the prior definition given in the appendix for two class problems (A.3.1). For the "standard" prior I used the prior definition that came with the sample classification problem provided with Radford Neal's software. These two prior definitions are summarized in table 2.

The results of these experiments are summarized in figures 10 through 14. Although both systems work well for the largest sample size ($N = 200$), the results for the "standard" prior degrade dramatically for the smaller sample sizes, while the results for the "uniform" prior remain good throughout all sample sizes.

Table 2: The prior definitions used for the results shown in figures 10 - 14

	"UNIFORM"	"STANDARD"
Input To Hidden Hyperprior Width	2.0	0.2
Input To Hidden Hyperprior Alpha	5.0	0.5
Hidden Bias Hyperprior Width	0.5	0.05
Hidden Bias Hyperprior Alpha	5.0	0.5
Hidden To Output Hyperprior Width	0.25	0.05
Hidden To Output Hyperprior Alpha	2.5	0.5
Output Bias Hyperprior Width	0.25	0.05
Output Bias Hyperprior Alpha	2.5	0.5

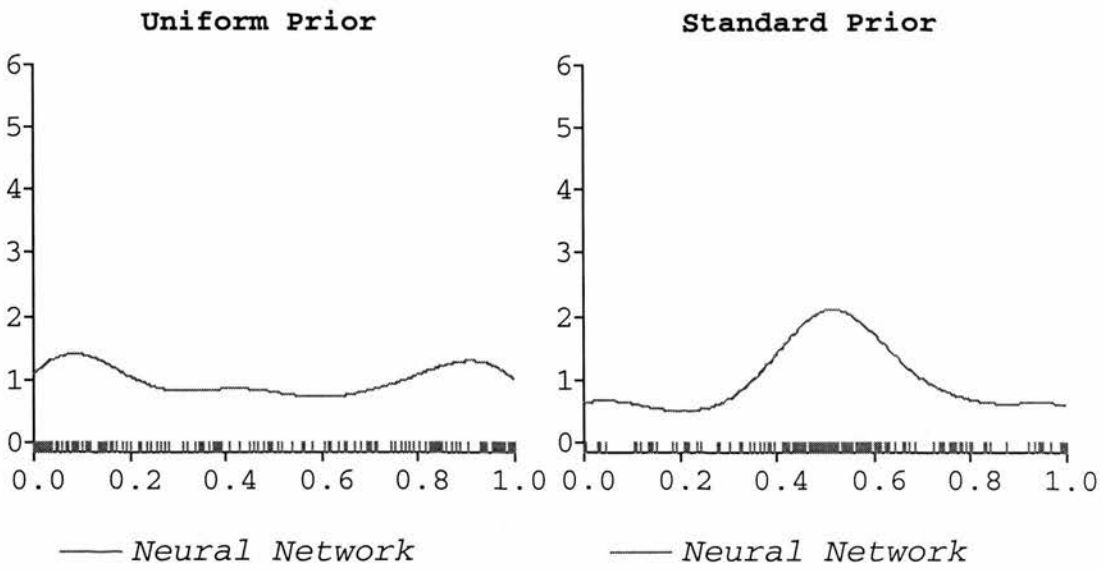


Figure 10: The two priors used in the following experiments

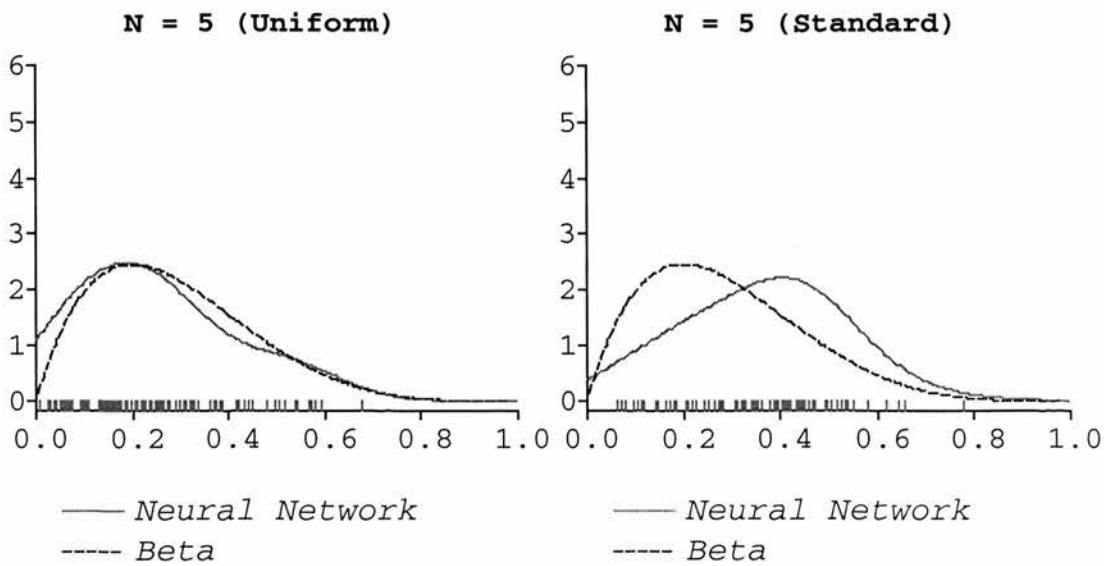


Figure 11: Uniform vs. Standard prior – Bernoulli trial, Sample size of 5. These results illustrate the importance of using a nearly uniform prior when dealing with sparse and noisy data, if the output distribution is to be interpreted as an approximation to a conditional probability distribution.

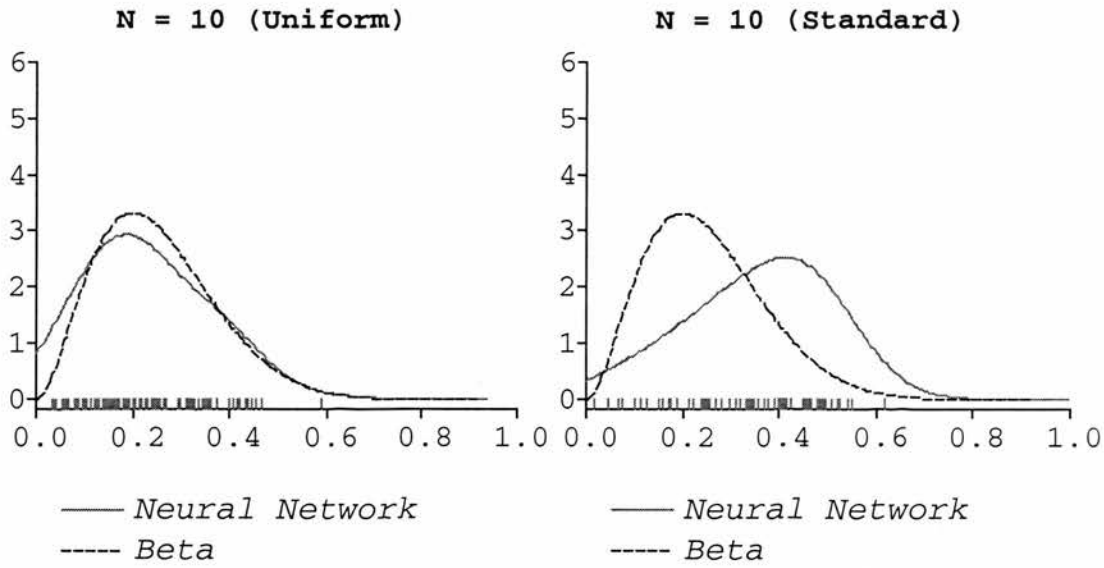


Figure 12: Uniform vs. Standard prior – Bernoulli trial, Sample size of 10

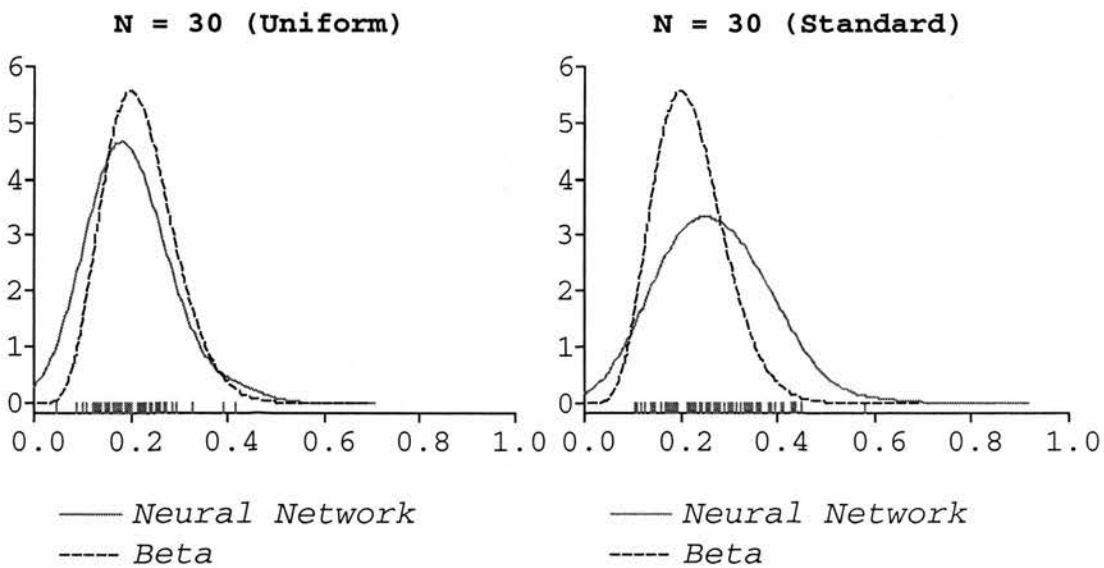


Figure 13: Uniform vs. Standard prior – Bernoulli trial, Sample size of 30

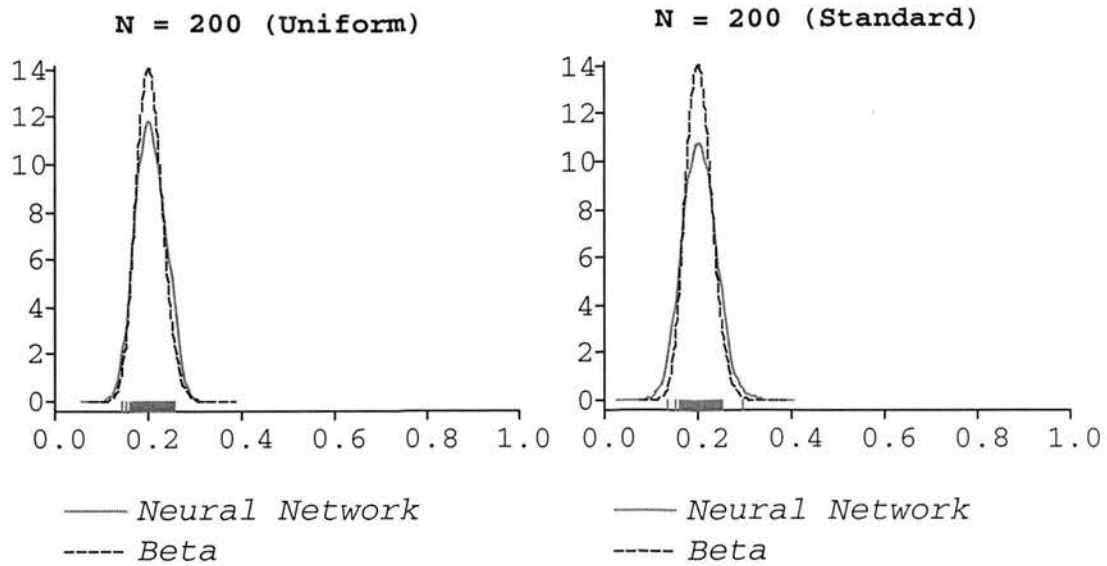


Figure 14: Uniform vs. Standard prior – Bernoulli trial, Sample Size = 200. The problems introduced by a poor prior can be overcome by a large data set.

2.9 The noisy XOR problem

In the previous example, the system was able to ignore its inputs and simply model the distribution of the target variable. Here I return to the noisy XOR problem introduced in section 2.3. To recap, I have generated data sets of various sizes by encoding the XOR problem, but for a fixed proportion of training cases for each of the four unique input vectors the target variable has been flipped to the opposite category. Thus we have finally come to a genuine categorization problem, albeit a very simple one.

Figures 15 through 18 display the output of the neural network system given exposures to this problem with varying training set sizes (*N*) and varying proportions (*P*) of "flipped" target values. I have continued to use for reference beta distributions computed for each unique input combination. That is, I compute these treating each possible input vector as an independent case. Since this uses my special knowledge

that the inputs are in fact independent, these should not now be viewed as the true probability distributions conditioned on the data. However, they do provide a frame of reference when inspecting the output distributions. Based on the results of a simulation using a training set with 800 examples (e.g. see figure 15), it appears that, given a large amount of data, the system is able to learn that the inputs are independent, and converges to a close approximation of the beta distribution.

As we decrease the training set size the approximations diverge from the beta distributions, especially for high noise levels. There is a consistent pattern. Relative to the beta distributions, the output distributions for these cases tend to be pushed in towards the center, and down towards a uniform distribution. This is due to the fact that the networks can no longer be as confident that the inputs are independent, so that the outputs for any particular input vector are to some extent influenced by the

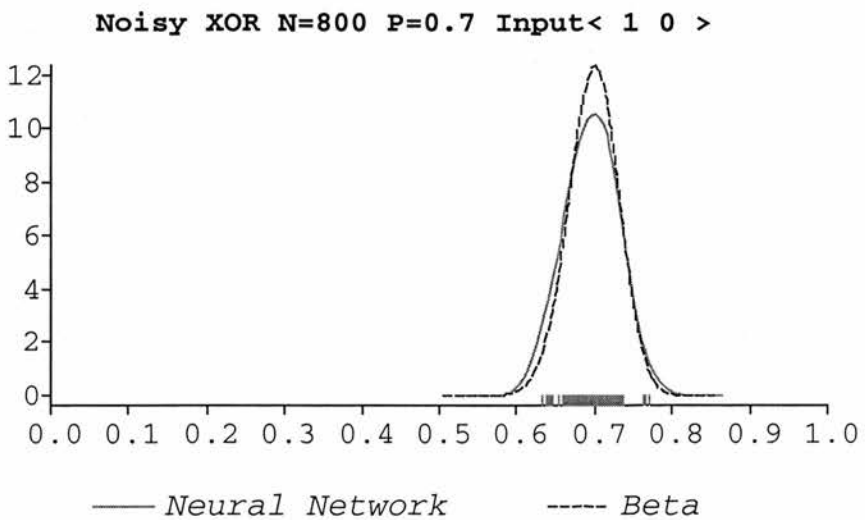


Figure 15 Noisy XOR predictions for the input (1, 0).: Training set size = 800 with the probability of the target value being set to 1 equal to 0.7

targets for the other vectors.. This behavior is desirable, since this is what permits generalization to inputs which have not been explicitly represented in the training set.

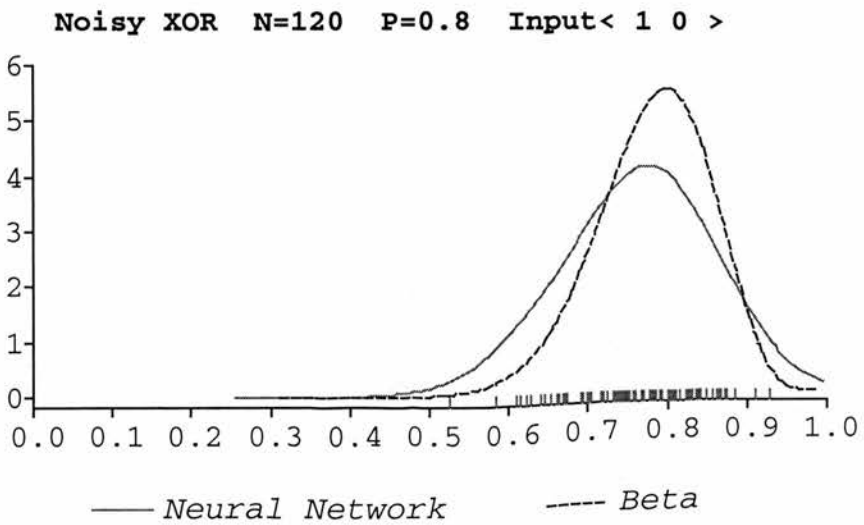


Figure 16 Noisy XOR Training set size = 120, P = 0.8, Input = (1, 0)

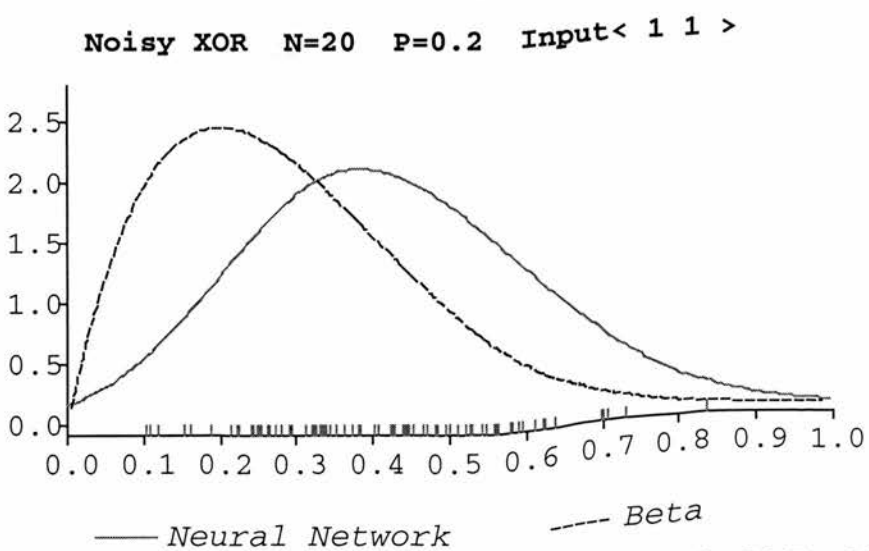


Figure 17 Noisy XOR:Predictions for input (1, 1) N = 20 P = 0.2

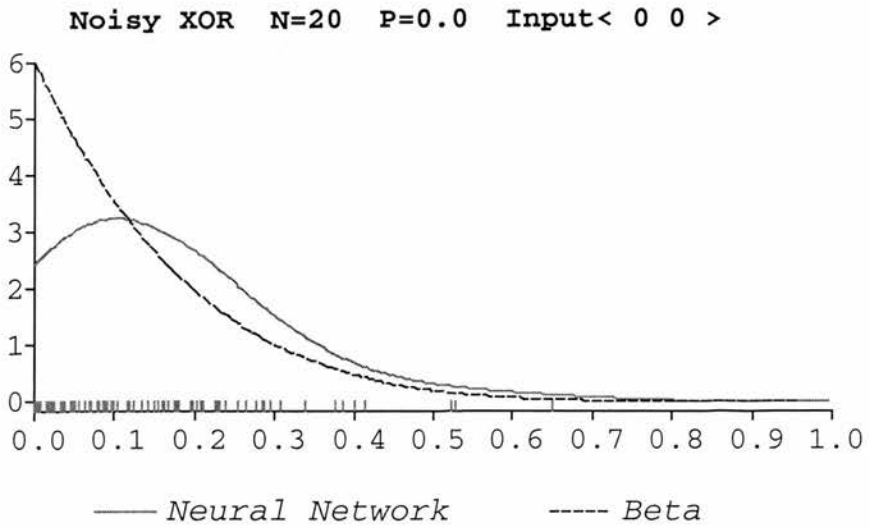


Figure 18 Noisy XOR predictions N = 20 P = 0.0 Input = (0, 0)

2.10 Discussion of the simulation results

I consider these results to be very encouraging. Neural networks are known to work very well as classifiers of complex nonlinear data when sufficient training data are available. This has led to the wide use of neural nets for problems such as speech recognition and optical character recognition. One criticism of the use of neural networks is that their performance is not easily interpretable. Another is that they may overfit given sparse data. Here I have applied a non-linear network with 41 adjustable parameters to a very simple problem, and the results are readily interpretable as approximations to conditional probability density functions. These accurately model the two kinds of uncertainty we are concerned with in a scientific application: uncertainty due to incomplete information, and uncertainty due to small sample size.

2.11 Probability Distributions over Three Outputs

In the previous section we looked at output distributions for two category problems. In our medical application, however, we may wish to model more than two possible outcomes. For example, in head injury research, the outcome categories used are often death, severe disability, and moderate disability to good outcome. A probability distribution over three categories lies in three space, but since probabilities sum to one, it is constrained to lie on the plane that passes through the points (1 0 0) (0 1 0) and (0 0 1). Since individual probabilities are constrained to the interval (0 1), these three points form a bounding triangle for the feasible region for probabilities in 3-space. A point in this triangle approaching one of the vertices represents near certainty for a particular outcome, while a point near the center represents a high degree of uncertainty. This triangular representation has been used extensively in modeling outcome following head-injury (Teasdale, 1981), and I will use it in the following sections

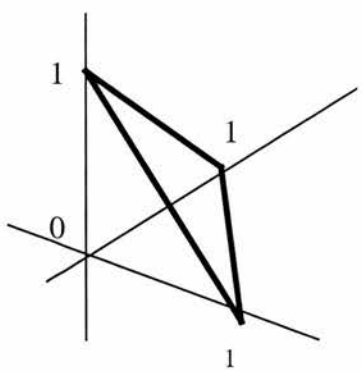


Figure 19: Feasible region for probability vectors in 3-space

2.12 Validating the three output model

To validate the three-outcome model, I repeated the "Bernoulli trial" experiment, but with three possible outcomes rather than two. I trained the network on a variety of probability distributions, which I kept constant over training sets of various lengths. Figure 20 shows the output of the system for Bernoulli trials of length 10, 40, and 120. Each of these are based on samples of 20% class 0, 50% class 1 and 30% class 2. Level contours are displayed for confidence regions of 90%, 70%, 50%, and 25%. As with the two class example, the continuous density is computed from the distribution of network outputs using kernel density estimation, and the true likelihood function is computed numerically. In this case the true conditional probability distribution is represented by a Dirichlet function rather than a beta function. The details of the computation of the Dirichlet function are given in the appendix (A.2.2).

2.13 Input Rescaling

Inputs to a neural network should be scaled so that they have similar mean values and standard deviation (Bishop, 1995). I have scaled input values using all available data for severely head-injured patients over the age of 14 in the Edinburgh database to determine the mean and standard deviation. At first I followed the usual procedure of rescaling the inputs so that they had zero mean, and unit standard deviation. In other words, I subtracted the mean value for a variable and divided the result by its standard deviation. I noticed problems with this procedure in applying networks to

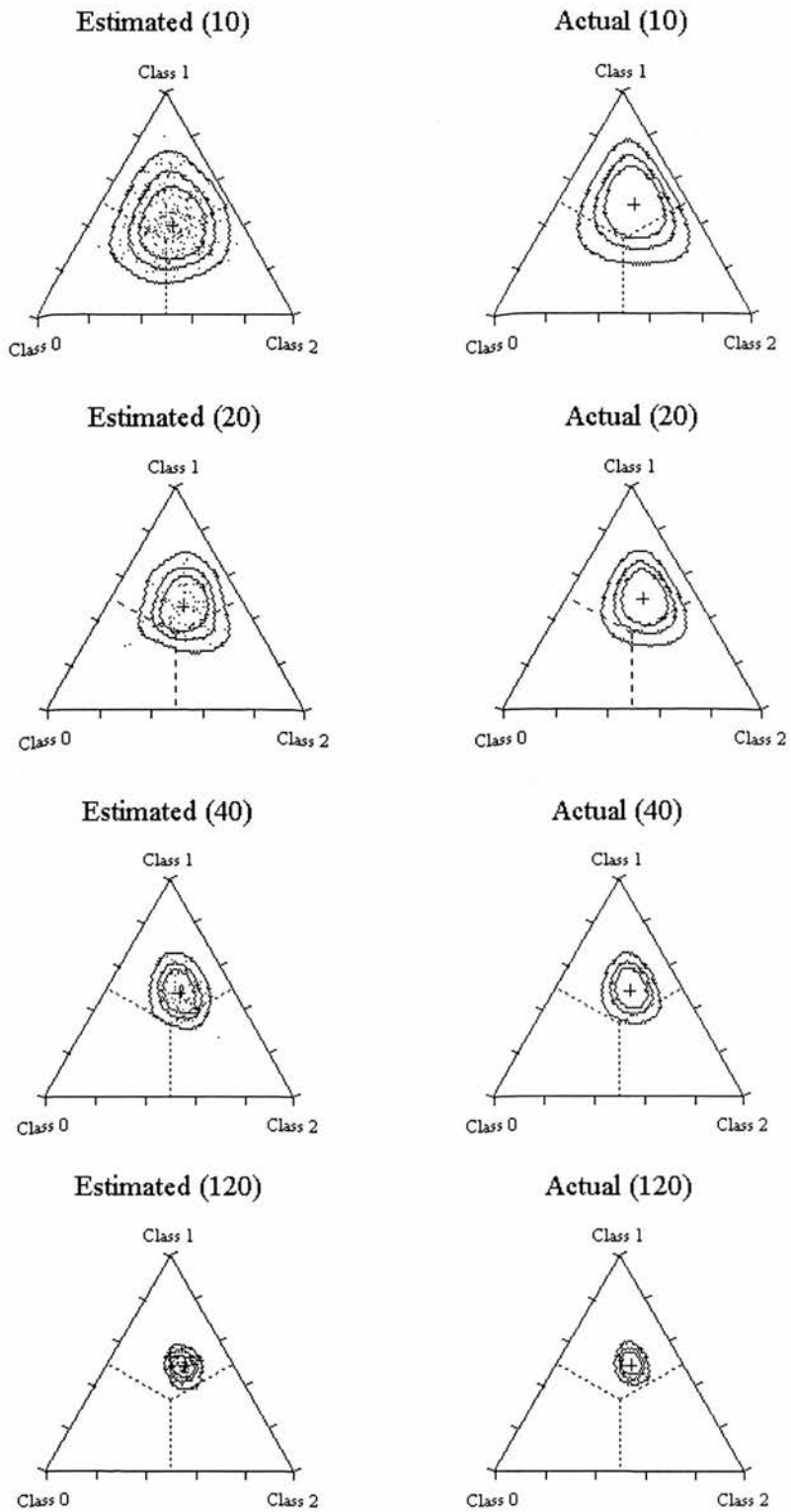


Figure 20 Actual and estimated densities with increasing sample size for the three output model. The "estimated" distributions are the outputs of the neural network, and the actual are the exact Dirichlet distributions computed numerically. The parenthesized numbers are the sample sizes.

sparse data, especially in regards to categorical and ordinal variables representing a small number of classes. Problems can arise when lopsided distributions over classes or values result in the mean being close to one extreme of the distribution. This causes the range of the variable on the other side of the mean to extend well beyond the central region of the sigmoid. To avoid this problem, I have centered the variable around the *midpoint* of its range, defined as the mean of the smallest and largest values seen, excluding values beyond 3 standard deviations from the mean.

2.14 Selecting the Prior

As discussed above for the purposes of this thesis, the prior distribution should be as neutral as possible. That is, in the absence of any data it should give rise to systems of networks that distribute predictions evenly over the space of possible probability vectors. I have found that it is important to calibrate the prior as precisely as possible. However, it's not possible for the prior to be perfectly consistent since a

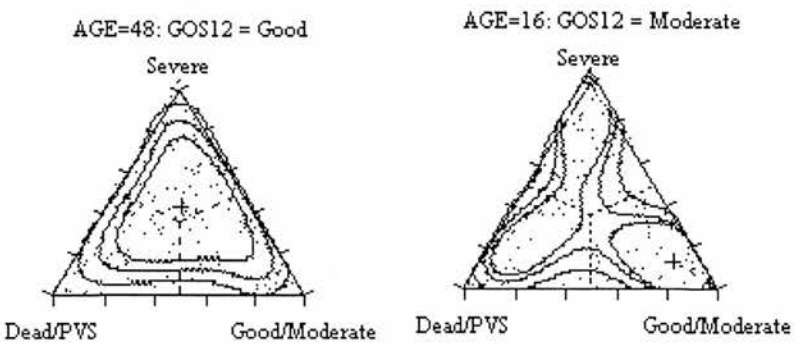


Figure 21 Outputs for 200 networks generated from the prior used in the following simulations given first a small magnitude input (left) and then a large magnitude input (right). This distribution is tri-modal, with peaks near each of the three vertices.

network generated by the prior will respond differently to an input with a small magnitude than to large magnitude inputs. I have found that the system performs best when the priors used slightly favor "strong" predictions (near the vertices of the triangle) when given large magnitude inputs, but slightly favor "weak" predictions (near the centroid of the triangle) when given small magnitude inputs (figure 21).

2.15 Generalization

So far I have used examples in which the input vectors are independent. That is, each input can be treated as a unique case with its own separate mapping to the probabilities expressed in the output values. An important assumption in using a model like a neural network is that this is not the case. Rather, we assume that the relationship between input and output values varies more or less smoothly, so that if you change the inputs only a little bit, then the outputs should change only a little bit. Exploiting this property, new instances can be classified by interpolating between cases in the training set.

A conveniently simple example is readily available in the Edinburgh head-injury database. One of the most important prognostic factors following brain trauma is pupil reaction. If both pupils react to light stimulation, then the patient will probably have a good recovery. If neither reacts the patient will probably die. If just one reacts, then the chances are about even. Figure 22 shows the results of an experiment I did to test the interpolation properties of this system. The right hand

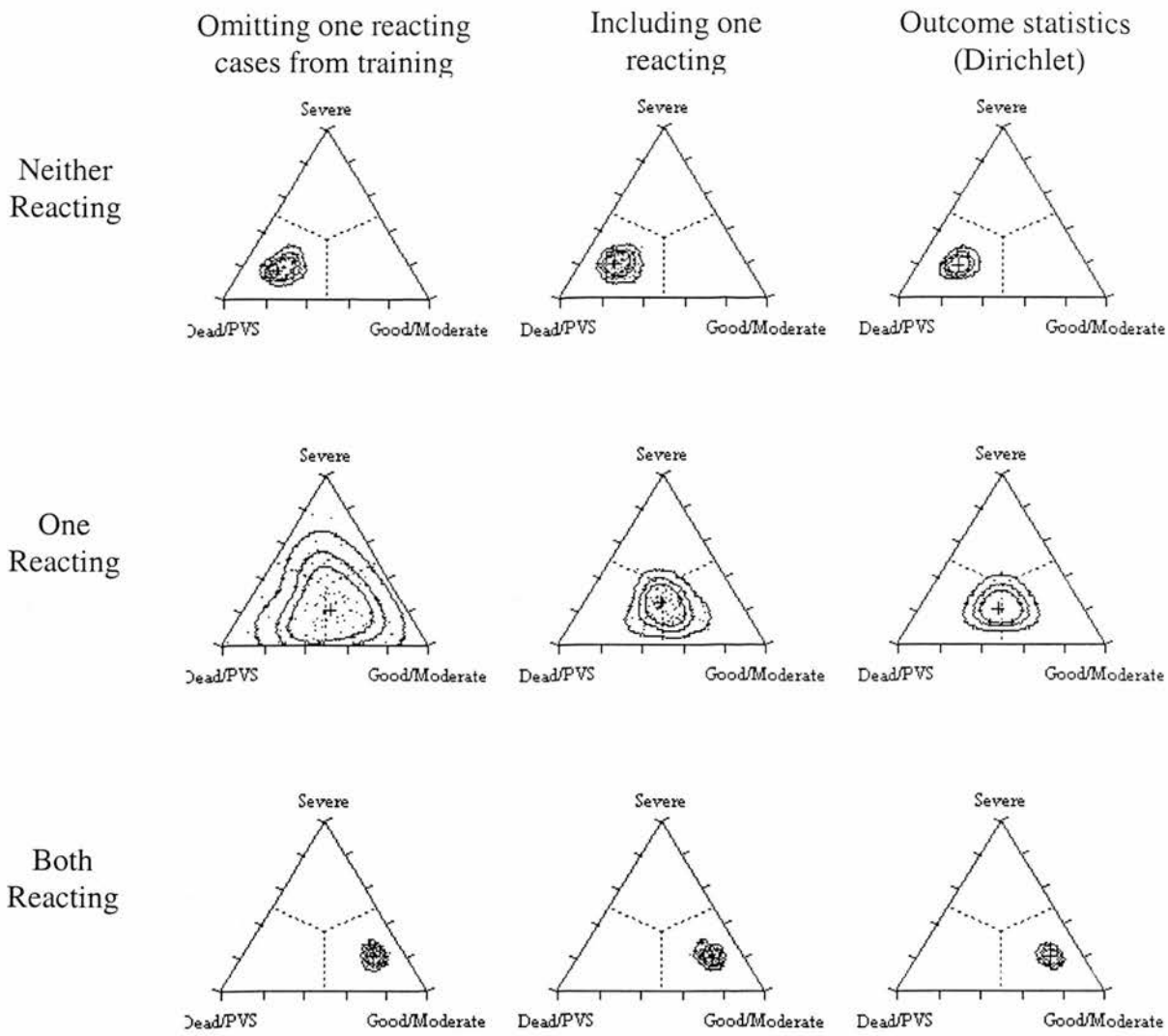


Figure 22 *Left Column*: outputs of networks trained on a dataset omitting all patients with exactly one pupil reacting on admission. *Middle column*: networks trained on the full dataset. *Right column*: the statistics of the database computed as Dirichlet distributions for each the three cases independently (see Appendix A.2.2). The input for the neural networks is coded as a single variable that can take on the value 0 (neither pupil reacting), 1 (one pupil reacting), or 2 (both pupils reacting).

column graphically displays the outcome distributions for each pupil reaction condition as represented in the Edinburgh database. These are Dirichlet distributions computed numerically as described in the appendix (A.2.2). The middle column

shows the neural network outputs for a system of networks trained on the database. The input in this training set consisted of a single variable that took on the value two if both pupils were reacting, one if only one pupil was reacting, and zero if neither pupil was reacting. The left-hand column shows the outputs for a system trained on the database omitting cases in which exactly one pupil was reacting. This forced the system to form its estimates for the case of one reacting pupil by interpolating between the two extreme cases. The mode of the predictive distribution is midway between the two other cases, which is where you would want an interpolating model to put it. The very loose confidence regions reflect the fact there is no guarantee that this guess is correct. The outcome categories are derived from the Glasgow Outcome Score (see chapter 1, table 3), with the two worst categories collapsed and the two best categories also collapsed, to form the three valued outcome set shown here.

As with the previous examples, the claim regarding the estimates formed by the neural network system is that they represent the conditional probability density functions for the output category probabilities conditioned on the data. Unlike most of the previous examples, it is not possible to compute the desired densities for comparison. However, the theoretical basis for this claim was summarized in the previous chapter, and the previous simulations for simpler problems have verified the accuracy of this implementation by comparison of the computed densities with the actual PDFs computed numerically. It is also reassuring that in this case modeling pupil reaction data the output densities for the full model correspond closely to the densities calculated by considering the three categories independently. It is not unreasonable to treat such a small number of ordinal categories as independent cases.

As was the case for two class problem, the numbers of hidden nodes used in the network did not appear to be a sensitive parameter in the model. The results in figure 22 were obtained with 14 hidden nodes. I tried varying the number from 4 to 20, and obtained similar results in each case. As discussed in chapter one, this property of the Bayesian neural network is of great significance. In other approaches to neural network training, the problem of determining the number of hidden nodes *is* critical, and usually requires the use of a special data set reserved just for this purpose.

The next chapter will proceed through a series of more complex examples using real data. This will allow us to further validate the Bayesian neural network implementation by comparing our results with well established results based on statistical prognostic modeling of patient outcome following head injury.

2.16 Summary

Historically, neural networks have proven to be successful engineering solutions to problems which are poorly understood, but for which large quantities of training data are available. Examples are the fields of speech recognition and optical character recognition. In scientific contexts, however, the usefulness of neural networks has been limited. This is partly due to lack of interpretability, and partly because the amount of available data for analysis would not usually be considered sufficient to justify the use of a highly complex model. Here I have applied networks with large numbers of hidden nodes to very simple problems, and the results are readily interpretable as approximations to conditional probability density functions. The

PDF's scale appropriately with sample size, and noise levels. This supports the argument that it is not necessary to avoid complex models in dealing with small datasets when this Bayesian framework is employed (Neal, 1996).

The examples in this chapter validate the concept of using Bayesian neural networks to support scientific research, and also verify the software implementation. Two key issues concerning setting the networks up for this kind of work are input normalization and specification of the prior on network parameters. Procedures for handling these problems have been implemented and tested. Interpolation properties of the system have been demonstrated on data from the Edinburgh head-injury database. Further tests of generalization and issues such as under- and over-fitting will be explored in the next chapters

The Edinburgh head-injury database consists of many overlapping datasets. Some of these consist of large numbers of patients, each described by a small number of variables. Others are composed of small numbers of patients, but each patient is described in terms of a much richer parameter set. The challenge in analyzing this database is to identify the most significant parameters being measured: those parameters that have the most to tell us about the pathophysiological processes that follow brain trauma. The two most important factors to consider are first, the conditional probabilities of various possible outcomes for each patient given the available data, and secondly, how secure those estimates are given the sample size for the particular parameter set being used. The modeling techniques described here provide principled descriptions of these two kinds of uncertainty.

Chapter 3

Prognostic models based on demographic data and simple clinical indicators

In the previous chapter I validated Bayesian neural networks on a series of very simple problems for which exact solutions could be derived. Now I will continue the model validation process on real data from the Edinburgh head-injury database. Comparison of results with exact solutions will no longer be possible. However, I will show that the use of Bayesian neural networks in modeling this data replicates results that are well established in the field of head injury research. This provides further validation of these new techniques. In the previous chapter I looked at a model based simply on pupil reactivity. Next, I'll move on to two slightly more complex variables. The first of these is the motor score component of the Glasgow Coma Scale. This will let us look at interpolation in Bayesian neural networks in slightly more complex contexts, moving from pupils, a three category ordinal variable to motor score, a five category ordinal variable. Then we'll look at age as a prognostic factor. This takes on values from 14 to over 90, so it approximates a continuous variable. Finally I will compare the performance of a series of multivariate models.

3.1 Baseline models using data available on admission

There is a long history of statistical modelling of outcome following head-injury based on demographic data and basic clinical information (e.g. Titterington et al., 1981). My intention here is not to compete with this work, or to compare the

performance of neural networks with other statistical techniques which have been used. Rather, I want to develop baseline models based on established principles in this field that I can build on in developing models that incorporate the detailed physiological data that are the distinguishing factor of the Edinburgh database. These baseline models will serve as controls for the more complex models. For example, a certain physiological effect detected in the Intensive Care Unit may simply be symptomatic of injury severity and not carry any new information. In this case, a baseline model incorporating the appropriate injury severity score will perform as well as the more complex model that also uses the physiological data. They will also allow us to further validate the implementation of Bayesian neural networks by comparison of our results with well established results in this field.

3.2 The motor score model

The motor score is a component of the Glasgow Coma Score (see chapter 1, table 1). It is a graded scale from one to five that measures the integrity of the patient's motor reactions as an indication of depth of coma. The scores used here are recorded after the patient's state has been stabilised in neurosurgery (post-resuscitation). Figures 1 and 2 show the predictive distribution of the neural network system compared with the statistics of the database computed as independent events.

As with the pupils score example, the model does not diverge a great deal from the independent data model, which is reassuring. To the extent that it does, it tends to linearize the model. This is reasonable. The overall trend of the model is

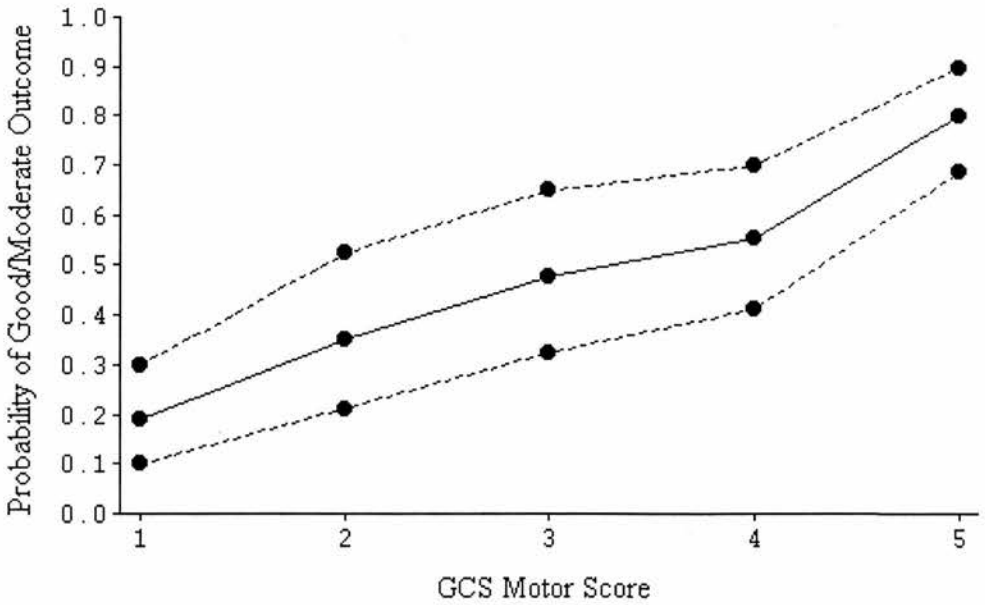


Figure 1 Probability of good/moderate outcome vs. motor score computed as independent categories for each motor score grade, with 90% confidence bands

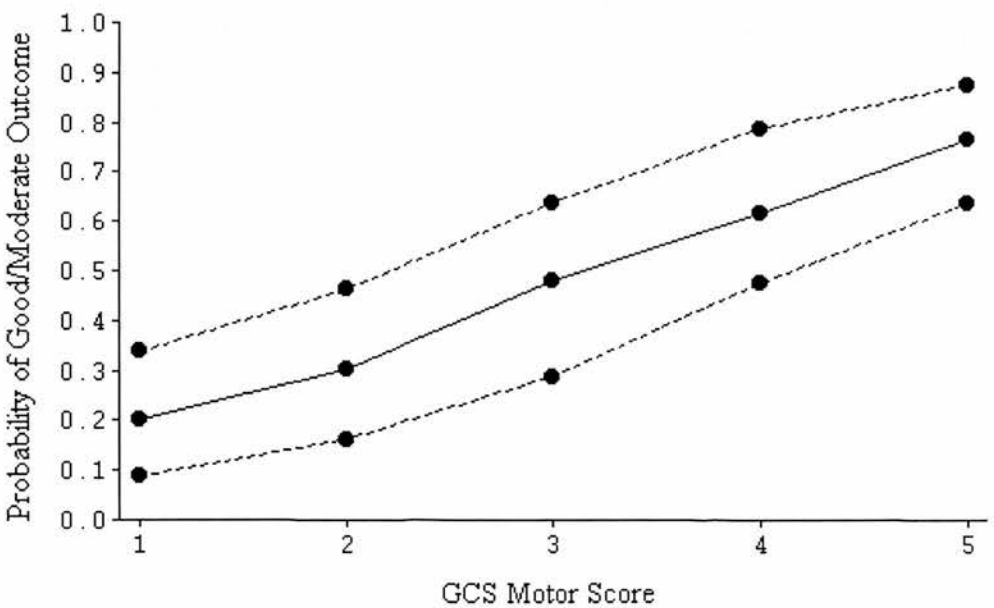


Figure 2 The neural network output distribution with 90% confidence bands

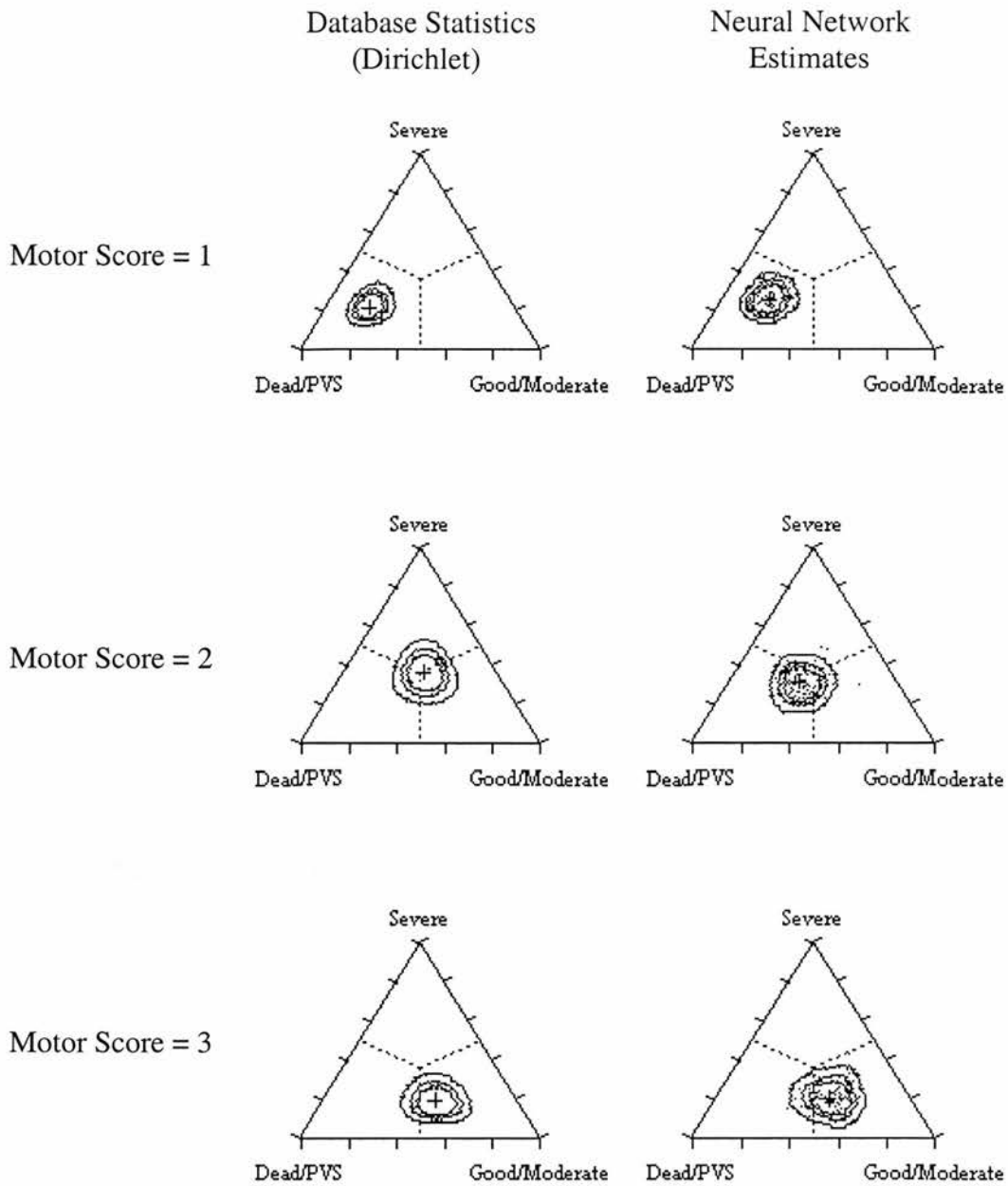


Figure 3 The predictions for outcome given motor scores 1-3 based on the statistics of the database considered independently (left) and on the outputs of the neural network system (right).

linear, so an interpolating model will tend to emphasize that trend due to effects of smoothing. The probability of good to moderate outcome given grade 2 motor score as estimated by the neural network is lower than the corresponding estimate using the independent data model. This seems slightly odd, since it appears to have the effect of making the relationship *less* linear. One should keep in mind however, that this plot is of only one of three probabilities, so some information is lost. A closer look shows why this has happened. Figure 3 shows how the full distributions change as they move from grade 1 to grade 3. Here it can be seen that the grade 2 score was much further outside the trend of the independent data model than was evident in the plot, and that the neural network model has pulled it more into line with the trend.

3.3 Modeling the effect of Age

Age is known to be a factor in recovery from brain trauma, with recovery being more problematic for older patients. The plot in figure 4 shows the probability of death against age as derived by a system of neural networks with twelve hidden nodes trained on age alone. In looking at the models trained on pupil score and motor score, I compared the outputs of the model to the raw statistics of the database. In this case, because the variable is much more fine-grained, such a comparison would be difficult to contrive and interpret. However, I have plotted the distribution of ages for patients who died, or were PVS at 12 months, as a rug of tic marks along the top of the graph. The distribution of ages for other patients is plotted as a rug along the bottom. In plotting the "rugs" I added a random fraction of a year onto the age to avoid an

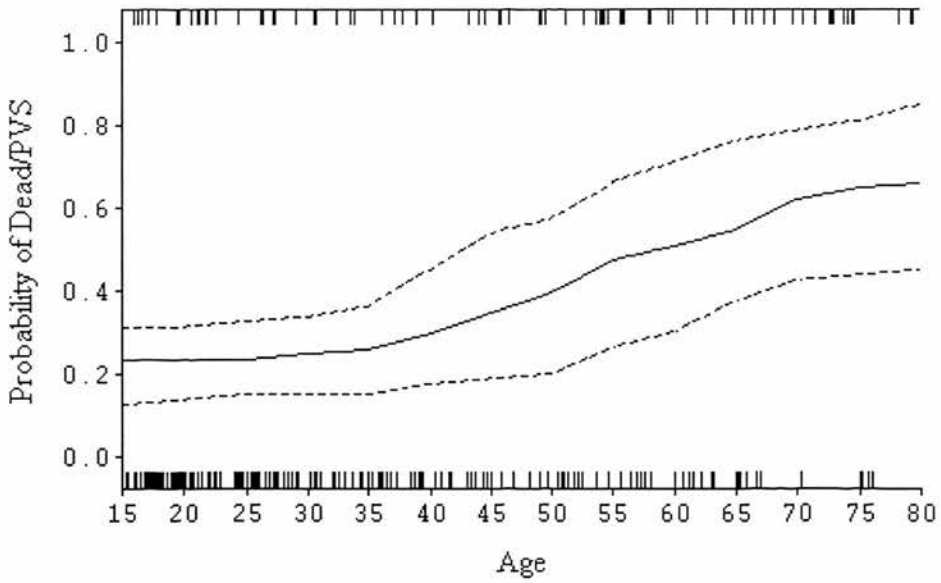


Figure 4 The effects of age on outcome following severe head-injury. This plots the neural network predictions for probability of death given age. The "rug" of tic marks along the top of the graph displays the distribution of ages for patients who died or were PVS. The lower "rug" is the distribution of ages for patients who did not die.

excessive number of collisions. The behaviour of this model replicates recent results working with the same database but using different techniques (Signorini, 1999a). That is, the effects of age are very slight until sometime in the 40's. Then the rate of increase of the probability of death is marked, and this trend continues in a roughly linear fashion into old age.

3.4 Comparing models

The three univariate models discussed so far, involving pupil score, motor score and age respectively, were discussed in terms of performance on their training sets. These models are sufficiently constrained that the functions computed can be inspected directly for smoothness and the effects of over and under-fitting. More

complex models are more difficult to evaluate. I hoped that performance on the training set might be an adequate basis for model comparison when Bayesian training techniques are used. However, when I compared results for a few models using three to five input variables first on the training set and then using 10-way cross validation, I found that predictive accuracy improved about 5 - 10% when evaluated in sample. I felt that this difference was too great to justify using the simpler approach, so on the remaining examples 10-way cross validation is used unless otherwise noted.

Two common measures for comparing the performance of models are the predictive accuracy measured as the percentage of correct test case predictions, or the difference between the percent correct and the score that would be achieved simply by guessing the most common class in all cases. These measures are very unstable given the small sample sizes we will be looking at. Often a few lucky or unlucky guesses near the decision boundaries make a model look much better or worse than it really is. Another problem with this metric is that it does not take into account the probabilities assigned by the model, but only the prediction itself. This is a particular problem for this study, because my primary interest is not in prediction per se, but in modelling the conditional probability structure of the data.

A more stable measure is the mean probability assigned to the correct class. However, this metric is not always a good indication of the degree to which the probability distributions have been correctly modelled. For example, say that the target variable varies randomly (with respect to the input variables) between two possible outcomes with a probability of 0.5. In this case a model that correctly estimated the probability of either outcome as 0.5 for all cases would be judged to be

no better than a model that alternated between the two outcomes, always assigning a probability of 1.0.

To avoid this problem we can work with the negative log of the probabilities rather than the probabilities themselves. This transformation produces a value that goes to zero when the probability assigned to the correct class goes to one, and diverges to infinity as the probability goes to zero. This value can be interpreted as our degree of "surprise" at a particular outcome given the assigned probability. I will favour the models that produce the least amount of "surprise" when the test data is revealed. The logarithmic score reported is the mean negative log probability assigned to the correct class. Again, I can correct for the effects of the arrival rates of the outcome classes by reporting the difference between the score attained by always guessing the probability distribution suggested by the arrival rates, and the score attained by the model. We will see that this score corresponds well to the degree to which the model is able to distinguish the outcome classes on the basis of the input variables.

I have calculated the output of the system differently depending on whether it is being evaluated according to percent error or logarithmic score. In the first case I have chosen the mode of the predictive distribution as the output of the system. As discussed in the previous chapter, this represents our best guess at the class of the test case given the available data. When the model is to be evaluated using the logarithmic score, I have integrated over the predictive distribution. This can be accomplished simply by taking the mean of the predictions of all the networks selected from the posterior distribution. This modifies the estimate of the probabilities assigned to each class based on the available sample size, avoiding

overconfident predictions when there is insufficient data to support them. This only makes sense when you are interested in accurately modelling the conditional probability structure of the data; there is no point in modifying predictions based on sample size if you are only interested in whether or not the correct class is predicted.

3.5 Comparisons of some simple models

I trained a series of models using basic demographic data and clinical indicators. The input variables used are described in table 1. The output is the Glasgow Outcome Score assessed at 12 months (see table 2)

The choice of input variables was based on current clinical practice and on previous work in modelling this kind of data. One of the most powerful predictors of

Table 1 Input Variables

Pupil Reaction	Both, one, or neither pupil reacts to light stimulation
Motor Score	Graded scale from 1 to 5 measuring motor response: a component of the Glasgow Coma Scale
Age	Coded as an integer continuous variable
Sex	Binary Variable
ISS	Injury Severity Score graded scale from 1 to 75
Focal	Binary variable indicating whether the injury is focal or diffuse based on the CT scan.

Table 2 12 month outcome categories used. Categories 1 and 2 were collapsed into a single category, as were 4 and 5, to produce the three output model.

GLASGOW OUTCOME SCORE	
1	Death
2	Persistent Vegetative State
3	Severe Disability
4	Moderate Disability
5	Good Outcome

outcome is the Glasgow Coma Score (GCS), which is composed of three separate scales for verbal, motor, and eye response. In the years since the GCS was devised, evaluation of the eye and verbal responses has become problematic, because seriously injured patients usually arrive at the neurosurgical unit already paralysed and ventilated. This makes it impossible to assess the complete GCS score without first allowing the effects of the drugs to partially wear off and removing the tube down the throat of the patient (Marion and Carlier, 1994). Most units will not do this unless it is clinically indicated. In the Edinburgh database, *all* of the severe head injury cases are recorded as having the lowest possible eye score, and 75% have the lowest possible verbal score, while the remainder have the next lowest score. Given the problems in assessment I decided not to include either of these components as inputs to my models, and to rely entirely on the motor score.

Besides the GCS motor score, I have included the pupil score which records whether one pupil or both pupils fails to respond to light stimulation. This is well known to be a strong indication of a poor outcome. I have also used the age and sex of the patients as predictive variables. Beyond these, I have tried including two other indicators. The first is the Injury Severity Score (ISS). This is an overall score indicating the severity of injury suffered by the patient including but not limited to the head injury. I included this because I thought it might be necessary to control for injuries to the patient besides the head injury. The other indicator is a binary variable indicating whether the injury was focal or diffuse.

The performance of the models tested is shown in table 3. As is to be expected with sparse and noisy data, the assessment using percent error leads to some anomalies: notably that the model including only pupil score and age outperforms all

Table 3 Performance metrics for simple models. N is the number of patients in the training set. The univariate Motor and Pupils models are evaluated in sample; the remainder are evaluated using 10-way cross validation. The error deltas are the improvements over always simply guessing the most common class.

INPUT VARIABLES	N	PERCENT ERROR	PERCENT ERROR DELTA	LOG ERROR	LOG ERROR DELTA
Motor	257	0.470	+0.055	1.007	+0.029
Pupils	249	0.386	+0.133	0.919	+0.114
Pupils Motor	248	0.412	+0.088	0.890	+0.142
Pupils Age	248	0.351	+0.151	0.911	+0.122
Pupils Motor Age	242	0.364	+0.141	0.832	+0.200
Pupils Motor Age Sex	242	0.360	+0.145	0.836	+0.196
Pupils Motor Age Focal	232	0.366	+0.150	0.843	+0.199
Pupils Motor Age ISS	235	0.370	+0.146	0.850	+0.189
Pupils Motor Age ISS Focal	226	0.354	+0.173	0.858	+0.179
Pupils Motor Age Sex ISS Focal	226	0.373	+0.151	0.869	+0.190

other models. The results using logarithmic score are more consistent with previous work. As explained in the previous section, the logarithmic error is generally more reliable than percent error in evaluation models. Here the optimal model is judged to be the one using pupil score, motor score and age, while the pupils/age model falls from first place to having the lowest score of any of the multivariate models.

The use of sex, Injury Severity Score, and the focal/diffuse variable do not seem to help prediction on this data set, and perhaps leads to some slight overfitting, as evidenced by the decline in accuracy of these models as compared to the pupils/motor/age model. In the next two sections I will compare these results with other modelling approaches used on earlier head-injury databases, and also on this same database.

3.6 Comparison with previous work on another database

One of the largest and best known previous studies is the comparative study of discrimination techniques based on the "three country" head-injury database discussed in the introduction (Titterington et al., 1981). A formal evaluation is not possible because of the different data sets collected in different decades, among other reasons. Even if it were possible, a full comparison with this work would be difficult because the authors' evaluation utilises 4 different variable sets, 6 different error terms, and 19 different modelling techniques. On the other hand, even a very crude and informal comparison will be of some value in validating the models employed here. If they do not perform at least roughly as well as established statistical techniques, we will know that something is wrong. I'll limit myself to a brief comparison of the neural network pupils/motor/age model with the results reported for their variable set II, which is the closest match possible. Variable set II includes four variables: age and the three components of the GCS score. For reasons discussed above, I have only included the motor component of the GCS. However, I have included the pupil score in training this model. Using error rate as a guide, the neural network model fares poorly, tying for dead last among the 20 models. It does better using the more reliable and interesting (from my point of view) logarithmic error term. Evaluated in this way it comes in 12^{th}. For details of the error rates for all these models see table 4. For a discussion of the "percent error" and "log error" error terms, see section 3.4 in this chapter.

MODEL	MODEL TYPE	PERCENT ERROR	LOG ERROR
NORLIN2	Normal based discriminant	0.306	0.757
NORLIN1	Normal based discriminant	0.316	0.760
INDEP2	Independence (Unordered Categories)	0.340	0.762
LINLOG	Logistic regression	0.314	0.764
INDEP3	Independence (Unordered Categories)	0.338	0.771
INDEP1	Independence (Unordered Categories)	0.338	0.775
LANC1	Lancaster	0.298	0.808
LANC2	Lancaster	0.298	0.809
LANC3	Lancaster	0.296	0.818
LATCL1	Latent Class (Mixture)	0.328	0.819
LATCL2	Latent Class (Mixture)	0.310	0.822
Bayesian Neural Network		0.364	0.832
KEREX3	Kernel (Ordered Categories)	0.340	0.852
KERORD2	Kernel (Ordered Categories)	0.332	0.856
KERUN2	Kernel (Unordered Categories)	0.328	0.872
NORQUAD	Normal based discriminant	0.304	0.884
KEREX2	Kernel (Ordered Categories)	0.326	0.903
KERORD1	Kernel (Ordered Categories)	0.352	0.905
KERUN1	Kernel (Unordered Categories)	0.364	0.924
KEREX1	Discrete Kernel	0.334	0.953

Table 4 *NB this is only intended to supplement the discussion in the text. This is* ***not*** *a formal model comparison.* This shows the results using Bayesian neural networks (BNN) on the Edinburgh database together with those reported earlier by Titterington et.al. using 19 different models on the "Three Country" database. The training set for the results reported there contained 500 patients, and the models were evaluated on a test set of equal size. The training set for the BNN contained 242 patients, and it was evaluated using cross validation. The inputs for the models also differed (see text).

I would emphasise two points in arguing that these results are actually quite encouraging:

- The training set size for the neural network model was less than half the size for the other models
- For reasons described above, it is difficult now to fully evaluate the GCS, so the current data for these kinds of models is not as informative as that available to the previous study

One further reason for a degree of optimism is that the previous study found that the simplest models worked best (naïve Bayes followed by the normal based linear discriminant). Despite having less than half the amount of training data, the neural network actually outperformed, judging by logarithmic error, the models of comparable power. It is not surprising that simple models would perform well on this data set. The indicants used are clinical scales designed to be monotonic and roughly linear. It is possible that the more complex models will compare more favourably when the problem becomes more complex, i.e. when we start adding in the physiological data.

3.7 Other work on the Edinburgh database

Two previous studies have developed predictive models based on the demographic portion of the Edinburgh database. The first (Signorini et al., 1999b) developed a logistic regression model with input variables age, pupil score, GCS score (the total of the eye, motor and verbal scores), ISS score, and a binary variable indicating focal or diffuse injury. This study reports an error rate of 15%. It's not possible to compare this result with those reported here or in (Titterington et al., 1981), however, for two reasons. First, this model does not attempt to predict severe disability as a separate outcome, but only distinguishes between survival and death. The second factor contributing to the lower error rate is that this study includes patients with moderate as well as severe head injuries. This leads to an improved error rate because moderates almost always have a good recovery by the criteria used here. Signorini et al. did not report the percentage of survivors in their test set, but 75% of

the patients in the training set survived. The proportion in the test set is probably similar. Therefore an error rate of 25% would have been attained simply by always guessing the most common outcome category (survival). By contrast, in the data set used for the "pupils/age/motor" model I reported, limited to severely injured patients, only 49.5% of the patients belonged to the most common outcome category (Good Outcome or Moderate Disability). In sections 3.4 and 3.5 I discussed the use of "delta" error scores, defined as the difference between the error rate achieved by the model and that achieved by guessing the most common class. If we use this as our criterion, the neural networks did better than the logistic regression model. The delta error (i.e. the improvement in prediction over guessing) for the neural networks is 14%, while that for the regression model is 10%. Of course, since these were evaluated on different data sets, this is not to be taken as a formal model comparison.

The modelling choices in the paper by Signorini et al. are reasonable given the aims of that study, but these choices would be counterproductive to the work described in this thesis. In looking at the effects of physiological derangements, the distinction between severe disability and moderate disability to good outcome is of particular interest. Modelling adverse physiological effects in cases of moderate head injury would be of interest, but moderates are rarely subject to intensive physiological monitoring.

The other previous study that developed prognostic models based on the demographic data in the Edinburgh database investigated the automatic induction of decision trees. This approach has the great advantage of producing models that may have direct clinical interpretations. However, of the models produced based solely on the demographic data and clinical indications discussed in this chapter, none have

performed better than simply guessing that the patient will have a moderate to good outcome (Andrews, McQuatt, et al., 1999).

3.8 Model performance with increased sample size and increased information per sample

As described in the previous chapter, my main interest in developing these models is to identify clinical features that allow us to discriminate between the patients who are doing well, and those who are not. To display graphically how well a model is able to distinguish the output classes on the basis of the inputs, we can average over the test set the predictions for each output class and plot these "prototype" probability vectors for each class in the prediction triangle as in figure 5. Figure 6 shows how discrimination improves both with increased sample size and with additional input

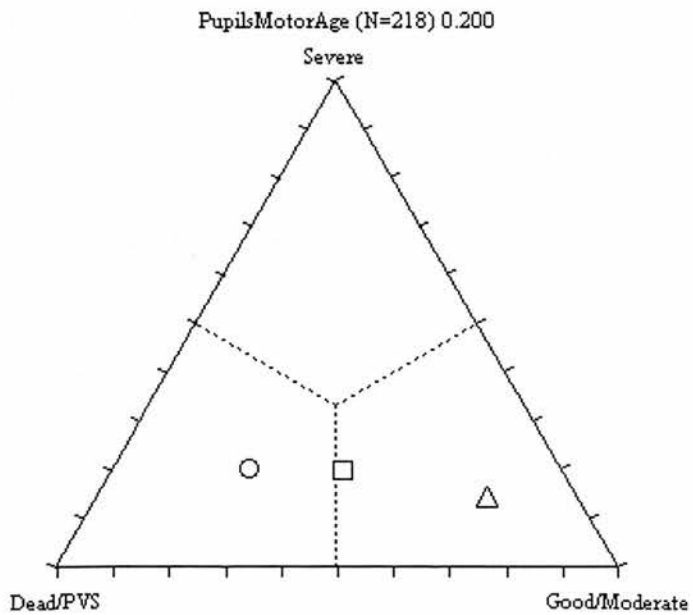


Figure 5 The mean of predictions made by a simple prognostic model for cases broken down by the three output classes: Death (circle), Severe Disability (square) and Good outcome (triangle).

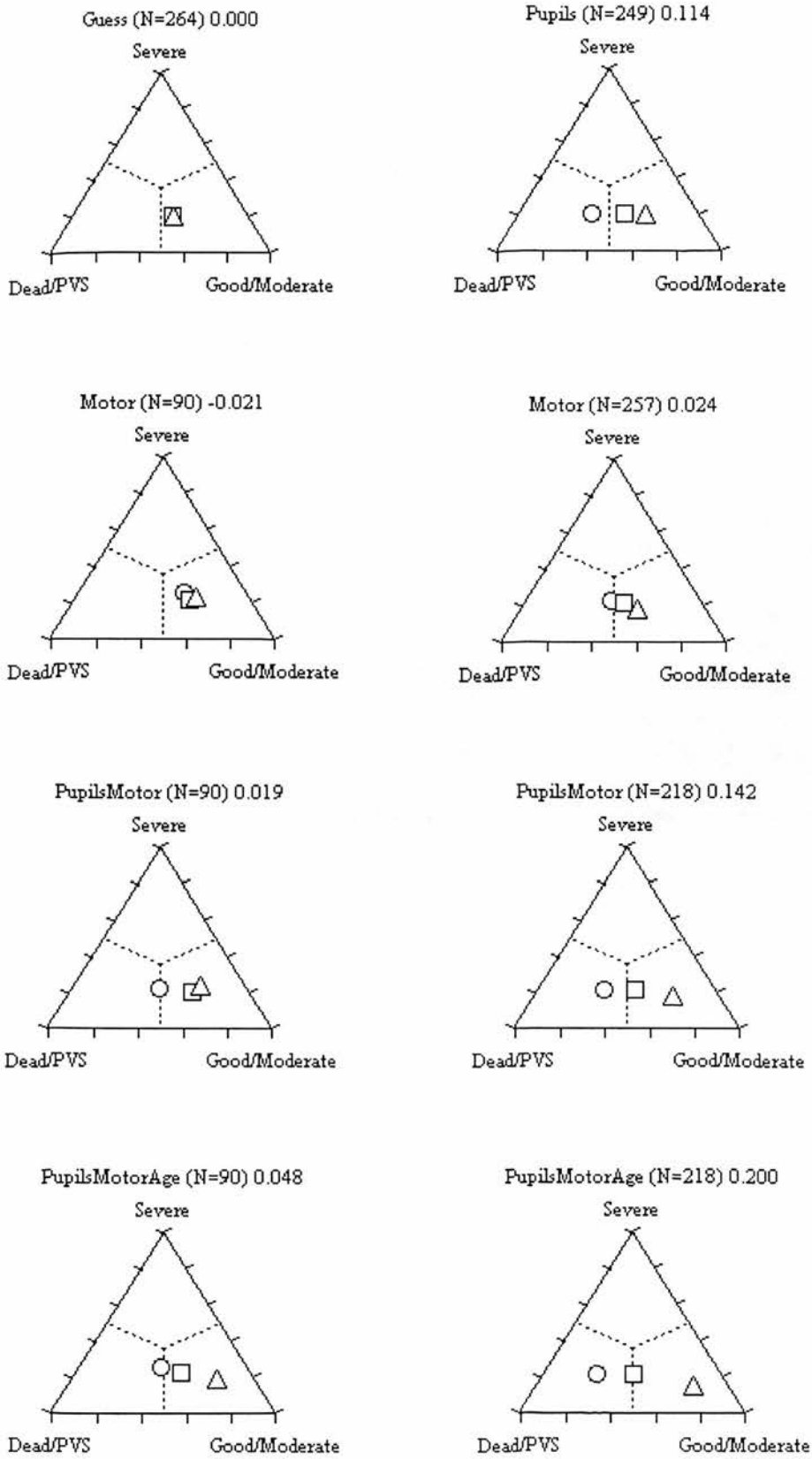


Figure 6 Increasing ability to separate classes with more input variables and increased sample size. Sample size is given in parentheses, the following number is the improvement in logarithmic score over guessing the prior probabilities. See table 1 for description of input variables.

variables for some of the models tested. The score shown for each model is the difference between its logarithmic score and the score attained by simply guessing the most common output class, i.e. good to moderate outcome. This appears to correspond well with the distance separating the prototype vectors in output space. It is important to note that this representation may appear to understate the clinical significance of the input variables. This is because the logic of these diagrams is the reverse of the usual course of clinical logic. For example, look at the motor score diagrams in figure 6. The icons representing the various possible outcomes are not well separated for the larger data set, and practically collapse for the smaller one. The logarithmic score for the small data set is actually worse than that attained by guessing. One might be tempted to conclude that that the motor score is not of great clinical significance. If we look at the plot in figure 2, however, we see a very different picture. Here we see the graph of the probability of a good outcome against motor score, and it's clear that as the motor score goes up, the probability of a good outcome goes up in a linear fashion. This plot depicts the normal relationship between clinical indication and patient outcome, answering the question: "Given certain indications, how likely is a particular outcome?" The plots I am using here on the other hand work in reverse fashion, answering the question: "Given that the patient had a particular outcome, how predictive of that outcome were the clinical indications?" These questions correspond to the probability of the outcome given the data and the probability of the data given the outcome. These two quantities are related mathematically by Bayes' theorem, but as this example shows, the relationship is not always obvious. This should be borne in mind when evaluating these and the following results. Where possible, for example in the case of univariate

models, I will use a representation like the one in figure 2, that plots outcome predictions against values on the input variable. Interpretation of multivariate models will be more difficult.

3.9 Summary

In this chapter I continued the process of model validation begun in the previous chapter. I first looked at a univariate model of the GCS motor score. This is a clinical scale that assumes integer values between one and five, with higher values being better. It was carefully designed to have a roughly linear relationship with outcome. The fact that this is a coarse grained variable allowed us to compare the neural network predictions with those obtained using the independent data model, and the results are reassuring. Next I looked at another univariate model, this one relating age to outcome. Since age is a much more fine grained variable, it was not possible to validate the model in the same way. However, the results obtained using Bayesian neural networks, replicate closely another analysis using this data and an additive model (Signorini et al., 1999b).

Finally I looked at a series of multivariate models relating standard clinical indicators to patient outcome. This allowed a rough comparison of the results using Bayesian neural networks with those reported in the literature using standard statistical techniques. The necessarily very informal nature of this comparison does not permit any conclusions regarding the relative merits of these different modelling techniques. The point of this comparison was to confirm that the performance of these neural networks on this part of the Edinburgh database was roughly similar to that obtained using standard techniques, and would lead to similar conclusions

regarding the relative importance of the standard clinical indicators. Since these results in the literature are very well established, any deviation would cast doubt on the validity of these new modelling techniques. On the other hand, the fact that these models produce results consistent with the previous work provides another measure of model validation.

At this point I will consider the model validation process to be complete. In the next three chapters I will apply Bayesian neural networks to the full Edinburgh head-injury database. The emphasis will be on the interpretation of the detailed physiological data, which is the distinguishing feature of this database.

Chapter 4

Feature extraction from physiological time series data

So far the models presented here of the data in the Edinburgh head-injury database have been based entirely on demographic data and simple clinical indicators. As noted in the previous chapter, there is a long history of prognostic modeling based on this kind of data. The work described to this point contributes little that is new, although it does validate the modeling techniques used, and it also provides some insights into this database. The distinguishing factor of the Edinburgh study is its focus on the influence of physiological derangements during intensive care on patient outcome, and on detailed computerized recording of patient physiology. This chapter will be concerned with methods for extracting features from this mass of data that can be used for prediction and analysis.

4.1 The Edinburgh University Secondary Insult Grades

The minute by minute recordings of physiological parameters measured in the intensive care unit have to be summarized before they can be presented to a prognostic model like a feed-forward neural network. I have been guided in my choice of features in part by the thresholds selected on clinical grounds for a previous study involving this database (Jones et al., 1994). A series of physiological insult severity thresholds were defined for each parameter being collected. These insult thresholds have collectively become known as the

Table 1 The Edinburgh University Secondary Insult Grades (EUSIG)

	GRADE A	GRADE 1	GRADE 2	GRADE 3
Raised Intracranial Pressure (mm Hg)	------	≥20	≥30	≥40
Arterial Hypotension (mm Hg)	------	≤70	≤55	≤40
Arterial Hypertension (mm Hg)	------	≥160	≥190	≥220
Cerebral Perfusion Pressure (mm Hg)	≤70	≤60	≤50	≤40
Hypoxia (SaO2 %)	------	≤90	≤85	≤80
Cerebral Oligemia (SvO2 %)	------	≤54	≤49	≤45
Cerebral Hyperemia (SvO2 %)	------	≥75	≥85	≥95
Pyrexia (°C)	------	≥38	≥39	≥40
Tachycardia (bpm)	------	≥120	≥135	≥150
Bradycardia (bpm)	------	≤50	≤40	≤30

Edinburgh University Secondary Insult Grades (EUSIG) and are summarized in table 1. In the previous study, the absolute durations of monitoring time over or under these insult thresholds were considered as prognostic factors. Hypotension, pyrexia, and hypoxia were found to be the most significant secondary insults for prediction of patient outcome.

A second study involving the Edinburgh database (Signorini et al, 999b), again using the EUSIG categories, tested four summary measures:

- Presence of any secondary insult for each category (binary variables)
- Total absolute duration for all secondary insults of the lowest grade or above in each category.
- Proportion of valid monitoring time spent at the lowest grade or above in each category
- A measure which weighted higher grades of insult more heavily than lower grades

For pragmatic reasons this study included all patients with computerized monitoring, and assumed that when a channel was not monitored (for example, due to the absence of an ICP probe), there were no insults of that type. This

study looked at three time periods, considering secondary insults during the first 24 hours, 48 and 72 hours following injury. For this reason, the distinction between absolute duration and proportion of good monitoring time is less significant than it otherwise would have been. All of the insult summaries performed equally well in terms of error rate. When evaluated using the Brier score (a quadratic error term) and ROC curve area, the weighted insult summary did best followed by proportion of good monitoring time. However, the results for these two approaches were so similar that the simpler feature using proportion of monitoring time *not* weighted by insult grade was selected for the final model. The results for all four feature extraction methods were that only ICP significantly improved prediction over a model based on simple demographic and clinical indicators.

4.2 A study based on the Traumatic Coma Data Bank

An earlier study based on the Traumatic Coma Database (Marmarou et al, 1991) systematically considered several different methods of feature extraction. For example mean, minimum and maximum parameter values during specified time periods were considered. In addition, absolute and proportional monitoring time were calculated for a wide range of candidate threshold values. In all this study entered 187 candidate descriptors into a stepwise forward variable selection procedure using a logistic regression model. It was reported that proportion of monitoring time with ICP over 20 mm Hg and proportion of monitoring time with arterial blood pressure under 80 were the two most prognostic factors of those considered. These values are close to the grade 1

insult thresholds used in the Edinburgh study, which were ICP over 20 and blood pressure under 70.

Due to the computational expense of the modeling techniques used in the present study, it has not been feasible to perform a systematic search of candidate features. I have relied on the previous work described above in limiting myself to proportion of monitoring time outside thresholds. The thresholds used have generally been the least severe grade insults defined in the EUSIG categories. In a few cases it was necessary to modify these thresholds in order to obtain a reasonable distribution over patients for a particular feature.

Using features based on threshold values has the advantage of simplicity, and is also consistent with clinical practice. Clinicians are accustomed to thinking in terms of trying to keep physiological parameters within certain limits. Arguments can be made for using either the absolute or proportional measures for insult duration. The results reported in (Signorini et al, 1998) favor the proportional representation, and this is consistent with results using a similar database collected at the Baylor College of Medicine (C. Contant, personal communication).

Finally, there is the question of how to deal with missing data. Signorini and colleagues (1999b) made the assumption that the absence of monitoring data for a particular parameter implied the absence of secondary insult. This allowed them to develop a prognostic model that can be used for any patient in intensive care, regardless of what monitoring data is or is not available. Since the focus of this study is on gaining a better understanding of the Edinburgh database rather than on prognostic modeling per se, I have adopted a different

approach. For any set of parameters I enter into a model, I require that each has at least 6 hours of valid monitoring time within 48 hours of the injury. This requirement limits the value of these models for actual prognosis in the ICU because they can't cope with missing data. On the other hand it removes a potential source of error in analyzing the existing body of data.

4.3 Summary

This study will build on previous work on feature extraction from physiological time series data collected from patients following a head injury. This work has consistently found that the percentage of monitoring time during which physiological parameters are below or above predetermined thresholds is a useful prognostic feature. It is also one that is easily understood from a clinical perspective. This is the approach I will adopt in this thesis, and my choice of thresholds will also be guided by this previous work.

Chapter 5

Raised intracranial pressure and related factors

Chapter two laid the groundwork for applying Bayesian neural networks in a scientific context. Chapters three and four surveyed previous work on prognostic modeling for head-injured patients, and further validated the neural network models by using them to replicate previous results in this field. In this chapter I will apply these models to the physiological time series data contained in the Edinburgh head-injury database. It is hoped that the use of flexible, nonlinear systems to model these detailed records of patient physiology will lead to new insights regarding the risk factors for head-injured patients in intensive care. In this chapter we will look at intracranial pressure, cerebral perfusion pressure, arterial blood pressure, and body temperature. It will be shown that for the patients studied here, the effects of all of these parameters on patient outcome are best understood in relation to a feedback loop between raised intracranial pressure and reduced cerebral perfusion pressure. This analysis will also provide insight into the critical threshold on cerebral perfusion pressure, and the evolution of these problems over time following the primary injury.

5.1 Intracranial pressure and cerebral perfusion pressure

Raised intracranial pressure (ICP) is probably the most studied of any of the physiological parameters discussed here. Because the brain is tightly enclosed in the skull, swelling due to contusions and inflammatory processes can rapidly lead to various forms of neuronal and structural brain damage, and in extreme cases, to

death. One problem associated with high ICP is a reduction in cerebral perfusion pressure (CPP), defined as the arterial pressure gradient between the cranium and the rest of the body. Increases in ICP can produce reductions in CPP, affecting the delivery of oxygen and nutrients to the brain. CPP cannot be measured directly, but

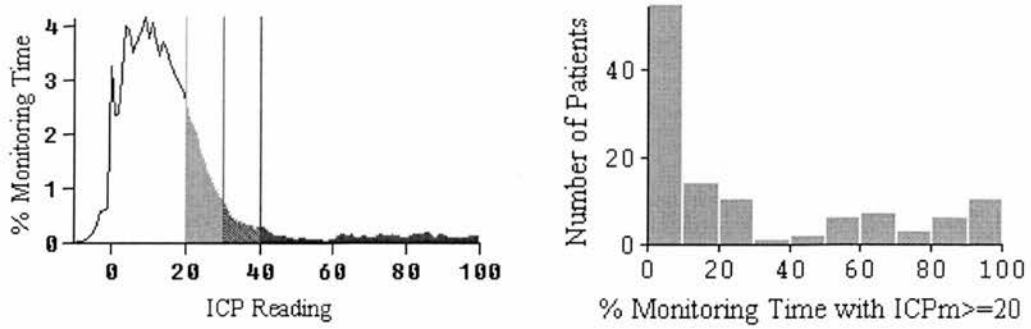


Figure 1 Left: the sampling distribution of ICP over the Edinburgh database. Vertical axis is percentage of monitoring time: horizontal axis is ICP in mm HG. The regions corresponding to the EUSIG grades are highlighted.

Right: the distribution over patients of the feature for proportion of monitoring time with ICP over 20 mmHg. Vertical axis is number of patients: horizontal is proportion of monitoring time spent with ICP above 20.

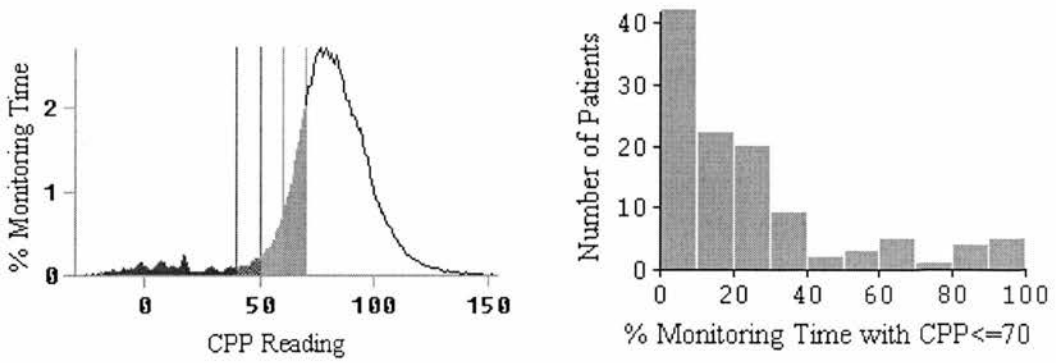


Figure 2 Left: sampling distribution for CPP readings over the Edinburgh database. Vertical axis is proportion of monitoring time: horizontal axis is level of cerebral perfusion pressure in mm Hg

Right: Distribution for proportion of monitoring time spent with CPP under 70 mmHg. Vertical axis is number of patients

it can be estimated as the difference between ICP and arterial blood pressure. Some research has suggested that the primary aim of therapy should be to maintain CPP rather than to reduce ICP (Rosner 1985, Miller et al., 1993). Figure 1 shows the distribution of ICP values recorded in the Edinburgh database. The most striking feature of this distribution is the persistent tail covering very high ICP levels. The histogram to the right shows the numbers of patients having ICP over 20 for the given proportions of monitoring time. Again there is a long tail into the high ranges, suggesting that there is a distinct group of patients with severe ICP problems. The distributions for CPP readings, and for proportion of monitoring time spent with CPP under 70 (figure 2), show a similar pattern.

Figures 3 and 4 show the predictions of neural network systems trained on these ICP and CPP features. The ICP system has proportion of monitoring time with ICP over 20 as its only input, and the CPP system uses proportion of monitoring time with CPP under 70. These plots tend to confirm the existence of two distinct patient groups in respect to these features. There is no discernible effect from either CPP or ICP insult until the proportion of monitoring time in insult range reaches about 40%. This threshold corresponds with the beginning of the long tails on the distributions for these parameters that we see in figures 1 and 2. Once the effect of ICP begins to be evident, it increases in roughly linear fashion with the amount of insult. CPP on the other hand, looks more like a step function. Risk of death increases dramatically beyond the 40% threshold, rapidly reaching that associated with the highest levels of ICP insult, and then plateaus at this level.

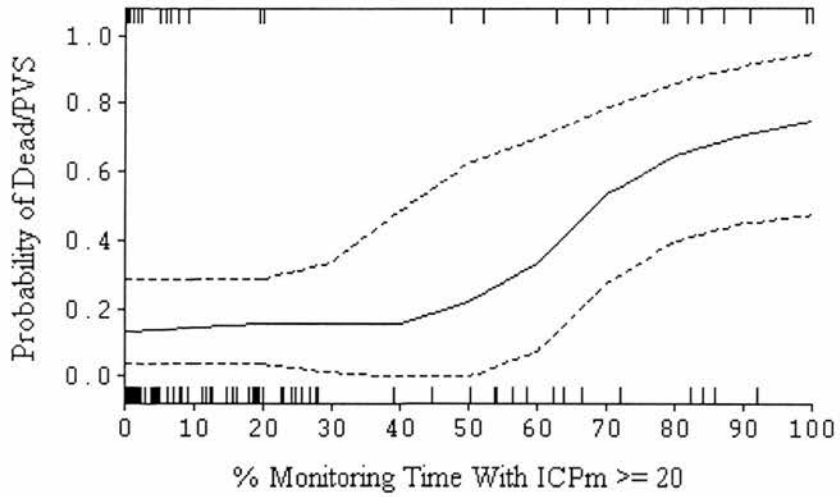


Figure 3 Neural network predictions of probability of death given proportion of monitoring time with ICP over 20. The dotted lines are 90% confidence intervals. The rug along the top is the distribution of this feature over the patients who died or were PVS. The rug along the bottom is the distribution of this feature for the other outcome groups.

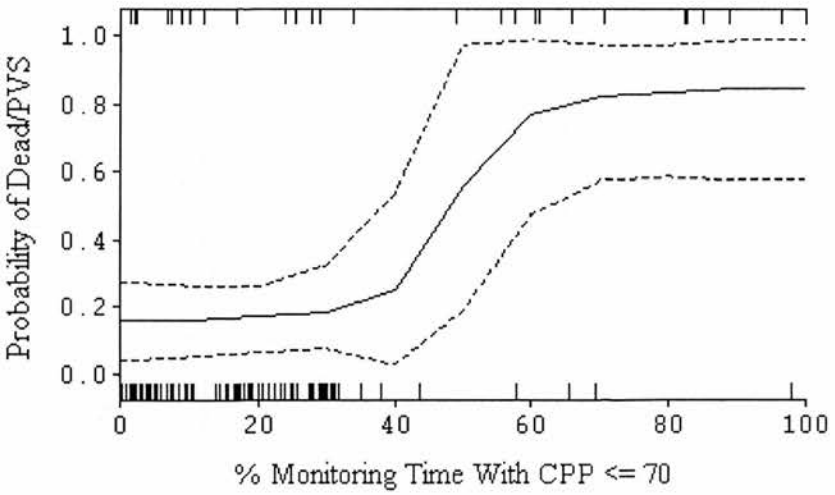
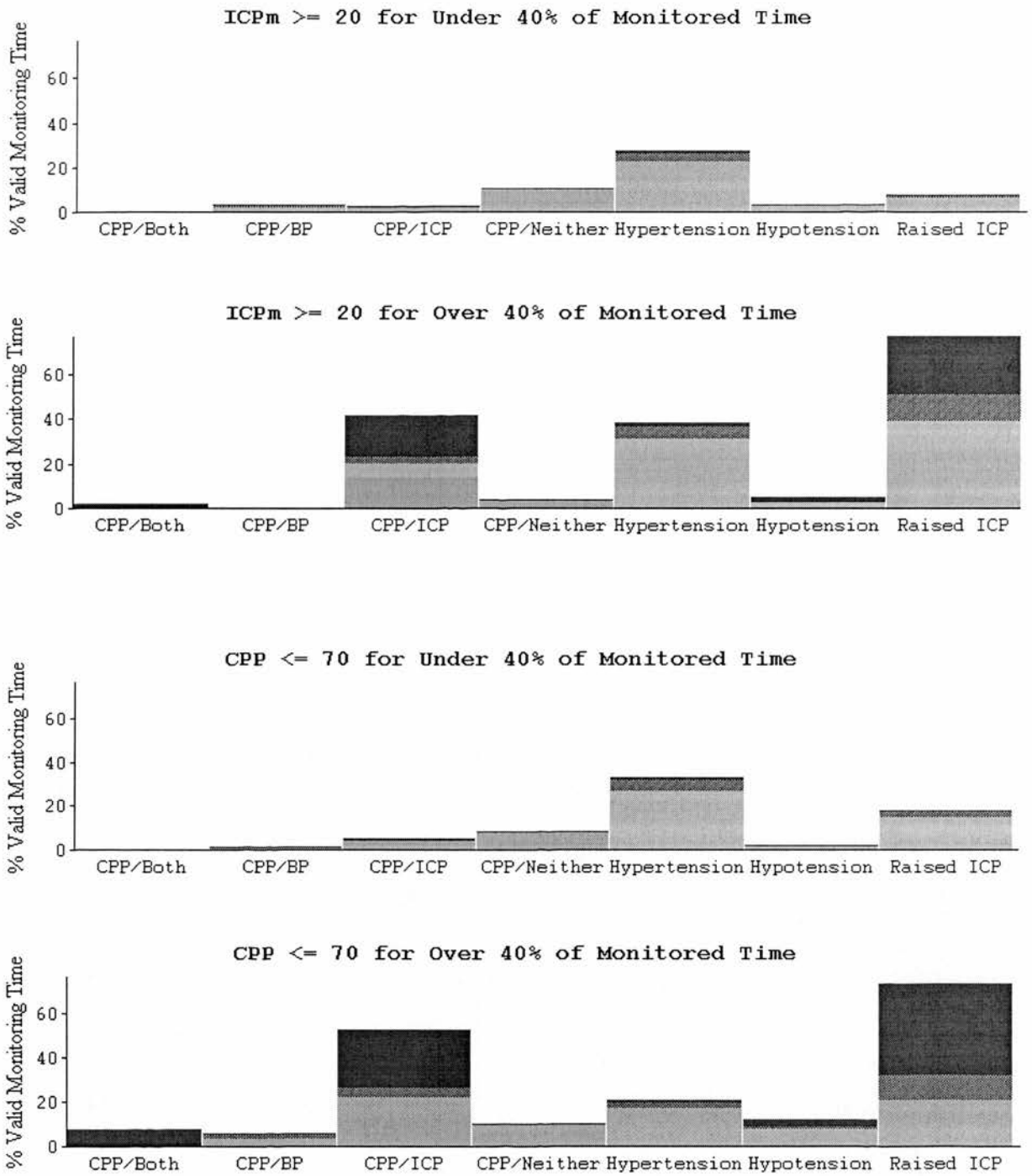


Figure 4 Neural network predictions of probability of death given proportion of monitoring time with CPP under 70.

I want to emphasise that the quality of the data does not justify viewing the 40% figure as an exact clinical threshold. The various patient histories and the amounts of missing data per patient make this figure difficult to interpret. Rather, it appears that we have two very distinct groups of patients. One group has relatively

small amounts of ICP/CPP insults, and the effects of these are relatively benign. Patients in the second group experience chronic problems, and these have severe consequences.

Of the 114 patients with ICP monitoring, there were 34 with ICP over 20 mm Hg at least 40% of the time, and 21 with CPP under 70 for at least 40% of the time. Figure 5 shows the distributions of CPP related insults for these groups and for the patients under the 40% thresholds. The bar plots are coded according to the grade of insult, with the most serious grades (CPP < 40 mm Hg and ICP > 40 mm Hg) being the darkest. Four kinds of CPP insults are broken down into those associated with ICP insults, those associated with arterial hypotension, those associated with simultaneous ICP and hypotensive insults, and those associated with neither of these. The most striking aspect of these plots is the difference between the under and over 40% groups in the most severe insult ranges. Almost the entire distribution of grade 3 insults is concentrated in the high risk groups. The overall insult duration is also much higher, with the exception of arterial hypertension. This is a special case since for most of these patients raised blood pressure represents an adaptive response that helps maintain acceptable levels of CPP. The dramatic increase of CPP insult in the high risk CPP group is due almost entirely to CPP associated with high ICP levels. Paradoxically, the most striking difference between the high risk CPP and the high risk ICP group is that the CPP group has a higher proportion of *very severe* ICP insults than the ICP group. This is because the CPP group excludes an intermediate range of patients who experience significant amounts of mild to moderate ICP insult, but who compensate with raised arterial blood pressure levels to maintain CPP.



- **Figure 5** The occurrence of CPP and related secondary insults as percentage of monitoring time in insult range. The thresholds used are the EUSIG grades. The darker shading indicates the more severe insult grades. The four graphs correspond to four patient groups defined in terms of the amount of CPP and ICP insult they experienced. The CPP insult is divided into four categories: CPP insult associated with raised ICP (CPP/ICP), CPP associated with arterial hypotension (CPP/BP), CPP associated with both of those other factors (CPP/Both), and CPP associated with neither of the other factors (CPP/Neither).

These patients are included in the "high risk" ICP group but in the "low risk" CPP group. The addition of these patients does not greatly affect the outcome statistics of the low risk group, which suggests that this is where they belong. However, their removal from the high risk group leaves it dominated by those patients with severe, unmanageable ICP and impaired CPP autoregulation: Hence the step function effect in the CPP predictions.

The three patient groups whose outcome distributions combine to create this effect are:

- Patients with ICP over 20 mm Hg less than 40% of the time. (Low risk ICP group, or Low ICP group)
- Patients with ICP over 20 mm HG more than 40% of the time but with CPP falling below 70 mm Hg less than 40% of the time. (High risk ICP/Low risk CPP group, or ICP Only group)
- Patients with ICP over 20 mm Hg more than 40% of the time and with CPP below 70 mm Hg more than 40% of the time. (High risk ICP/High risk CPP group, or ICP/CPP group)

The outcome distributions for these three groups are plotted in figures 6 through 8. These highlight the difference between the high ICP group with CPP maintained and the group in which CPP regulation has failed. It is clear from these distributions that raised ICP with CPP maintained is far less significant than raised ICP leading to a reduction in CPP. These results provide support to the view that CPP is a critical factor in ICP management. On the other hand, the strong association between high levels of CPP insult and intractable ICP raises the concern that the patient group identified as having significant amounts of CPP insult may simply be untreatable. One consideration in evaluating these results is that these patients were

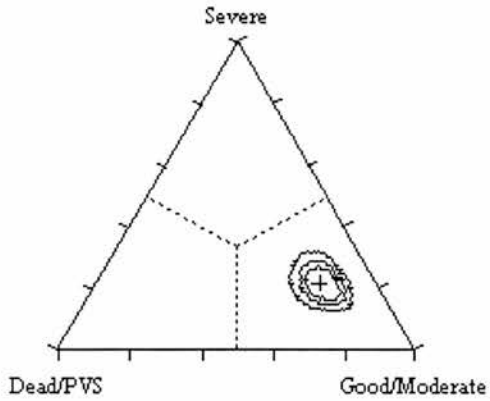


Figure 6 Outcome plot for the 80 patients in the low risk ICP group. This and the following two plots are actual outcome distributions from the database = *not* neural network predictions distributions.

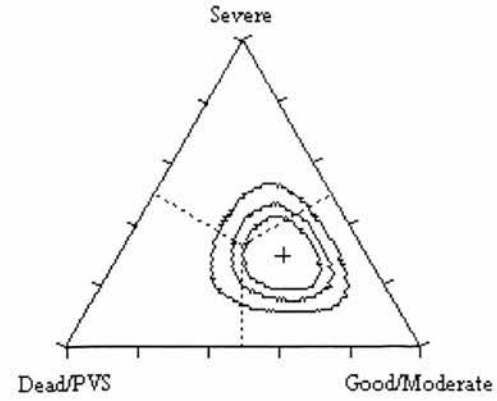


Figure 7 Outcome for the 17 patients in the high risk ICP but low risk CPP group.

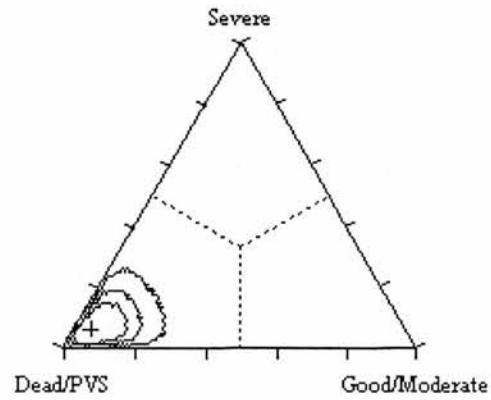


Figure 8 Outcome for the 17 patients in the high risk groups for both CPP and ICP

treated according to a protocol based on CPP maintenance. A comparison with results from centres using other protocols would be of interest. A look at the distributions of secondary insults for the high risk ICP/low risk CPP, and the high

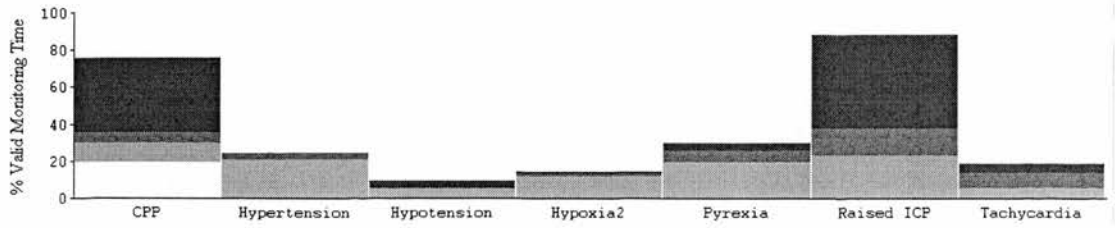


Figure 9 The associated secondary insults for patients in the high risk ICP/high risk CPP group.

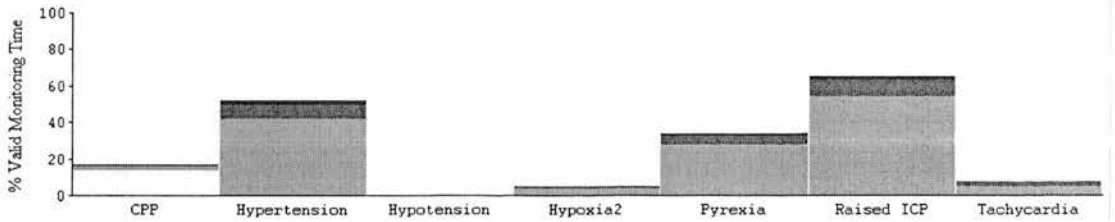


Figure 10 The associated secondary insults for the high risk ICP/low risk CPP group.

risk ICP/high risk CPP groups (figures 9 and 10) emphasises the difficulty in distinguishing between ICP and CPP problems in these patients because of the strong association between significant amounts of CPP insult, and very severely elevated ICP.

There is one group of patients that has not been discussed. These are the patients who have over 40% monitoring time with CPP under 70, but under 40% monitoring time with ICP over 20: i.e. the high risk CPP/low risk ICP group. This is a rare combination in the Edinburgh database; only four patients fall into this category. Two of these had a good outcome, one died, and one suffered severe disability. The data for these patients is summarised in table 1. Rather than treat these patients separately I have included them in the "low risk ICP" group. Because there are only four patients in this group, it is difficult to say very much about the

Table 1 Descriptive data for the four patients with significant amounts of CPP insult not strongly associated with raised ICP. The last column indicates whether the CPP was associated with ICP insults, hypotension insults, or neither of those. The last patient experienced reduced CPP along with some periods of mild ICP and hypotension.

AGE	SEX	CAUSE	REACTING PUPILS	MOTOR SCORE	OUTCOME	CPP PROBLEMS
17	Male	Pedestrian	Both	3	Good	Hypotension
33	Male	Assault	Both	5	Good	Hypotension
17	Female	Car Accident	Both	3	Severe	Neither
54	Male	Assault	Neither	1	Death	All Types

significance of reduced CPP when this is not accompanied with high ICP. However, it should be noted that while the combination of CPP and ICP problems has been shown to be a very strong predictor of death (88% mortality), of these four patients only one died and two had a good recovery. Therefore it is unlikely that reductions in CPP that are not accompanied by high ICP are nearly as dangerous as the two conditions combined.

5.2 Controlling for admission factors

In the previous section I distinguished three groups on the basis of ICP and CPP insult. The group with high amounts of both ICP and CPP insult has much higher mortality than the other two groups, which are similar to each other in terms of outcome distribution (figures 6, 7, 8). It is important to consider the possibility that this might have been predicted on the basis of their admission data alone. It has been

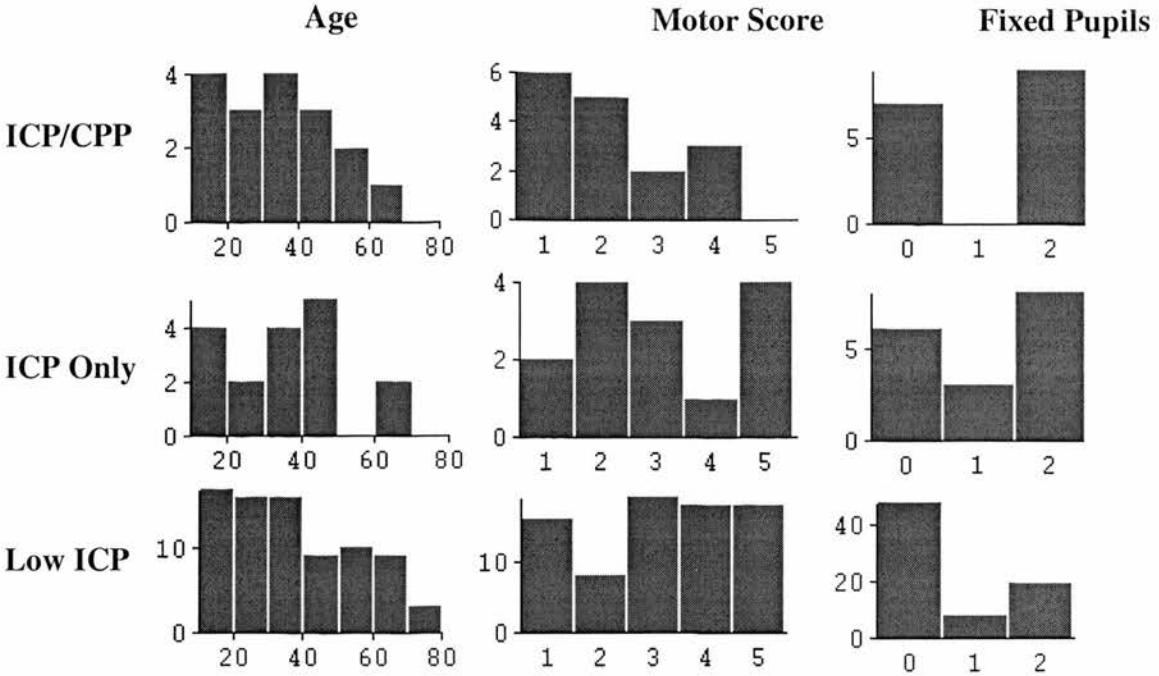


Figure 11 The distributions for age, GCS motor score, and pupils score for (from top to bottom) the high risk ICP/high risk CPP group, the high risk ICP/low risk CPP group, and the low risk ICP group as defined above.

Table 2 Summary of the key admission variables for these three patient groups. Age is summarised by the group mean. Motor and pupil scores are summarised by the group mean conditional probability of death or persistent vegetative state given that score. The calculation of these probabilities is described in the text below this table.

	GROUP MEAN	**CONDITIONAL P(DEATH/PVS) (GROUP MEAN)**				**ACTUAL MORTALITY**	
		Given Motor Scores		Given Pupil Scores			
	Age	P(death)	Std.Err.	P(death)	Std.Err.	Mortality	Std.Err.
ICP/CPP	35	0.31	0.07	0.42	0.04	0.88	0.08
ICP Only	35	0.27	0.06	0.42	0.05	0.24	0.10
Low ICP	37	0.27	0.06	0.30	0.04	0.15	0.04

previously shown that recovery following head injury can in part be explained on the basis of a few simple clinical variables that are available on admission, and on the age of the patient. It may be that the patients who suffered from a combination of

ICP and CPP insults were simply the ones who were in the worst condition when admitted to the intensive care unit.

Figure 11 shows the distributions for the key prognostic variables available on admission for the three risk groups. Table 2 summarises the significance of these variables in respect to the three patient groups. By inspection of the distributions and the group means there does not seem to be much difference between the three groups in terms of age. It is possible to summarise the coarse grained variables, motor score and pupil score, in terms of the group mean probability of death or PVS given the score. These probabilities are based on the proportions of deaths or PVS outcomes for the different values of these variables in the Edinburgh database, as summarised in tables 3 and 4, and on the distributions of these scores in the three patient groups (tables 5 and 6). Referring to the first two rows of table 2, except for the mortality rates, there is very little difference between the high risk group (ICP/CPP) and the group with significant amounts of ICP but not CPP insult (ICP Only). An assessment based only on this admission data significantly understates the risk of death for the high risk group, and overstates the risk for the other two groups. This

Table 3 The probability of death given that the patient has zero, one, or two fixed pupils based on the statistics of the Edinburgh database.

FIXED PUPILS	TOTAL CASES	DEATH/PVS CASES	ESTIMATED P(DEATH/PVS)	STANDARD ERROR
2	81	51	0.630	0.054
1	24	9	0.375	0.099
0	146	23	0.158	0.030

Table 4 The probability of death given Glasgow motor score based on the statistics of the Edinburgh database.

MOTOR SCORE	TOTAL CASES	DEATH/PVS CASES	ESTIMATED P(DEATH/PVS)	STANDARD ERROR
1	69	23	0.333	0.057
2	39	12	0.308	0.074
3	41	14	0.341	0.074
4	50	13	0.260	0.062
5	58	7	0.121	0.043

Table 5 The distribution of Glasgow motors scores for the high risk ICP/high risk CPP group, the high risk ICP/low risk CPP group, and the low risk ICP group.

	Motor Score					
	1	**2**	**3**	**4**	**5**	**Unknown**
ICP/CPP	6	5	2	3	0	1
ICP Only	2	6	3	1	4	1
Low ICP	16	8	19	18	18	1

Table 6 The distribution of numbers of fixed pupils for the high risk ICP/high risk CPP group, the high risk ICP/low risk CPP group, and the low risk ICP group.

	FIXED PUPILS			
	0	**1**	**2**	**Unknown**
ICP/CPP	7	0	9	1
ICP Only	6	3	8	0
Low ICP	48	8	19	5

provides support to the view that continuous ICP and CPP monitoring plays a critical role in the management of these patients.

5.3 The time course of ICP and CPP insults

Figure 12 shows the development over time of ICP and CPP insults for the three patient groups I defined above. The top row displays data for a subset of 43 patients selected from the "low risk ICP" group because they had at least one hour of

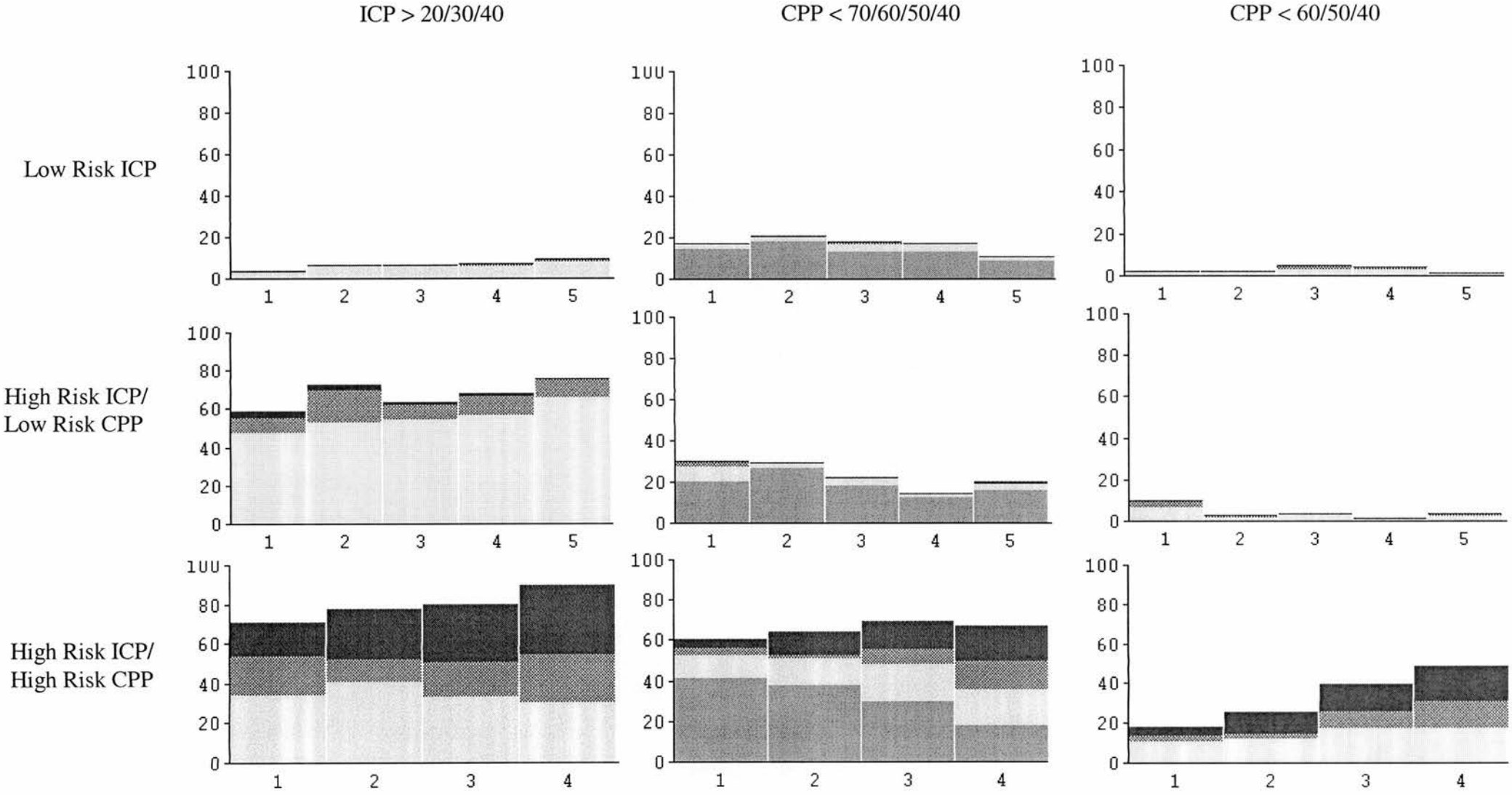


Figure 12 The time course of ICP and CPP insults for three patient groups plotted over the first five days for the first two patient groups, and over the first four days for the third group CPP insult has been plotted twice: with and without lowest category of insult (70- 60 mm Hg).

valid monitoring time in each of the first five 24 hour periods starting with time of injury. The middle row represents 14 patients in the "high risk ICP/low risk CPP" group selected in the same way. The bottom row represents 10 patients from the "high risk ICP/high risk CPP" group who had at least one hour of valid monitoring time in each of the first four 24 hour periods following time of injury. I stopped analysis at this point because adding in the fifth day would have further reduced this group to only eight patients. The first column shows the progress of ICP insults calculated as proportion of monitoring time over 20, 30 and 40 mm Hg. All three groups show an increase in ICP insults throughout this period. The second column in figure 12 displays CPP insult, plotting proportion of monitoring time spent with CPP in insult range using the thresholds of 70, 60, 50 and 40 mm Hg. More severe insults are more darkly shaded. The two groups that show little or no effect from reduced CPP have substantial amounts of insult. Over 80% of this is in the lowest range, between 60 and 70 mm Hg. Another anomaly involving this lowest insult range concerns the progress of the third patient group: the group that *is* at risk from CPP insult. We know from the fact that 8 of these 10 patients died that their conditions were generally deteriorating very seriously. However, the amount of the lowest grade insult suffered by these patients steadily declines. Between days 3 and 4 the decline in the amount of the lowest grade insult actually leads to a decline in the total amount of CPP insult reported for these patients.

These observations seem inconsistent with the idea that CPP between 60 and 70 mm Hg is harmful. Setting the CPP threshold this high leads to reporting significant amounts of insult for patients who do not seem to be being harmed, and to reporting relative reductions in the amount of insult being suffered by patients who

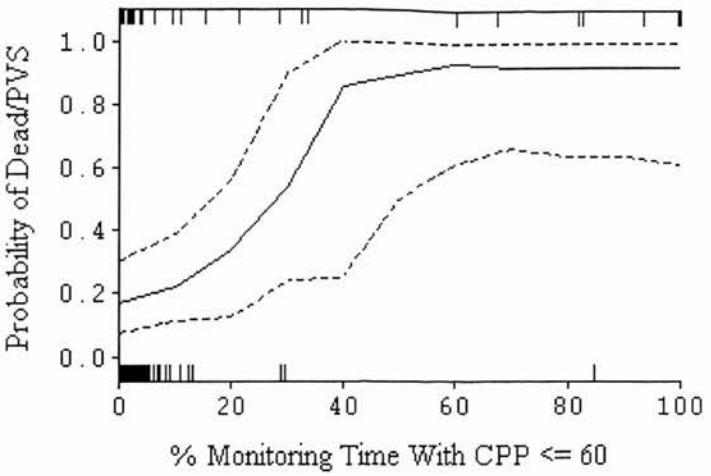


Figure13 The probability of death given proportion of monitoring time with CPP under 60 mm Hg.

are surely being harmed, and whose conditions are deteriorating. A more accurate picture of the effects of reduced CPP is seen in the third column of figure 12. Here we can clearly see the absence of CPP insult in the first two patient groups, and also the severity of the deterioration being suffered by the third group. We can retrain the neural network system using percentage of monitoring time with CPP less than 60 as the input feature rather than basing it on the higher threshold. The result is shown in figure 13. If we compare this with figure 4, there are two striking differences. First, we now see an immediate response to any amount of insult. Secondly, the response is now more graded. In figure 4 almost all of the change in probability of death occurs as proportion of monitoring time in insult range goes from 0.4 to 0.6. Using the 60 mm Hg threshold, the increase in probability of death is spread over the range of 0.0 to 0.4. Hence the clinically significant range over which we see a relationship between this feature and outcome is about twice as wide as that for the feature based on the higher threshold. Also the range on the Y axis (probability of death) has been

increased, indicating a greater degree of influence on outcome, and a better separation of classes. The range using the 70 mm Hg threshold is from 0.16 to 0.85, while for the 60 mm Hg threshold, the range is from 0.17 to 0.93. These results show that, at least for these patients, the 60 mm Hg threshold is more relevant clinically and produces a more robust prognostic feature than the threshold of 70 mm Hg.

5.4 Arterial Hypotension

Since CPP is determined as the difference between ICP and arterial blood pressure, we can treat arterial hypotension both as an independent predictor of patient outcome and as a component of reduced CPP, following an analysis similar to the treatment of ICP in the previous section.. I have used proportion of monitoring time with arterial pressure under 80 mm Hg as the feature representing hypotension. The 80 mm Hg threshold is the one selected as the most significant in respect to patient outcome in the study by Marmarou and colleagues (1991). Figure 14 shows the distribution of this parameter over the patients in the Edinburgh database. This is based on 118 patients who had at least 6 hours of valid monitoring time within 48 hours of the time of injury. Figure 15 plots the neural network predictions for the three outcome categories as proportion of monitoring time under threshold goes from 0 to 1. There is a step up in probability of death at about 30% of monitoring time under threshold. We can divide these patients into a high risk group (more than 30% of monitoring time with blood pressure under 80 mm Hg) and a low risk group (less than 30% of monitoring time under 80 mm Hg). Figures 16 and 17 show the

associated secondary insults for these two groups. The patients in the group with large amounts of hypotensive insult also experience large amounts of very severe CPP and ICP insult while those in the low risk group are almost completely free from these problems. This suggests that we should subdivide the high risk group into those patients at high risk from ICP and CPP insult and those at risk from hypotension alone. In the previous section I defined a group of patients at high risk from both ICP and CPP as those having over 40% of monitoring time with ICP over 20 and over 40% of monitoring time with CPP under 70 mm Hg. There are four patients in the high risk hypotension group who are also in the high risk ICP/CPP group. There are then 12 remaining patients in the high risk hypotension group, and 100 in the low risk hypotension group. Two patients are excluded because ICP data was not collected. The four patients at high risk for ICP, CPP and hypotension were all very severely injured; all had bilateral fixed pupils post-resuscitation. All four of

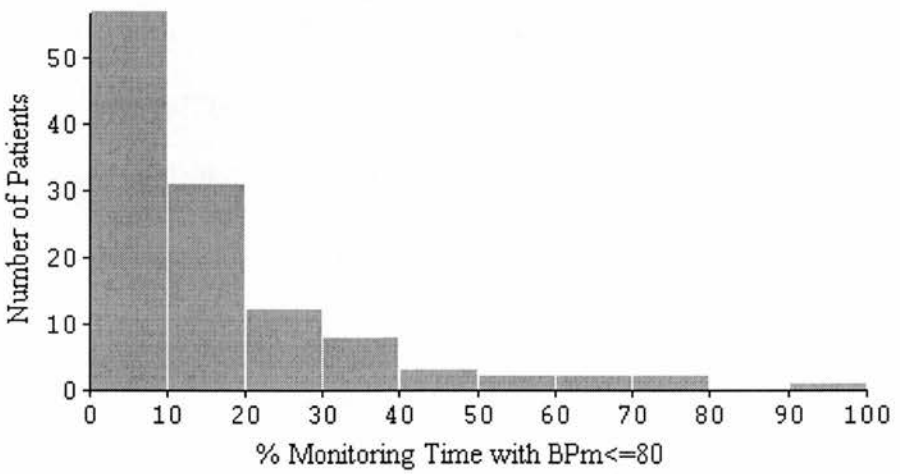


Figure 14 the distribution of the feature representing proportion of monitoring time with arterial blood pressure under 80 mm Hg. The vertical axis is numbers of patients, and the horizontal proportion of monitoring time.

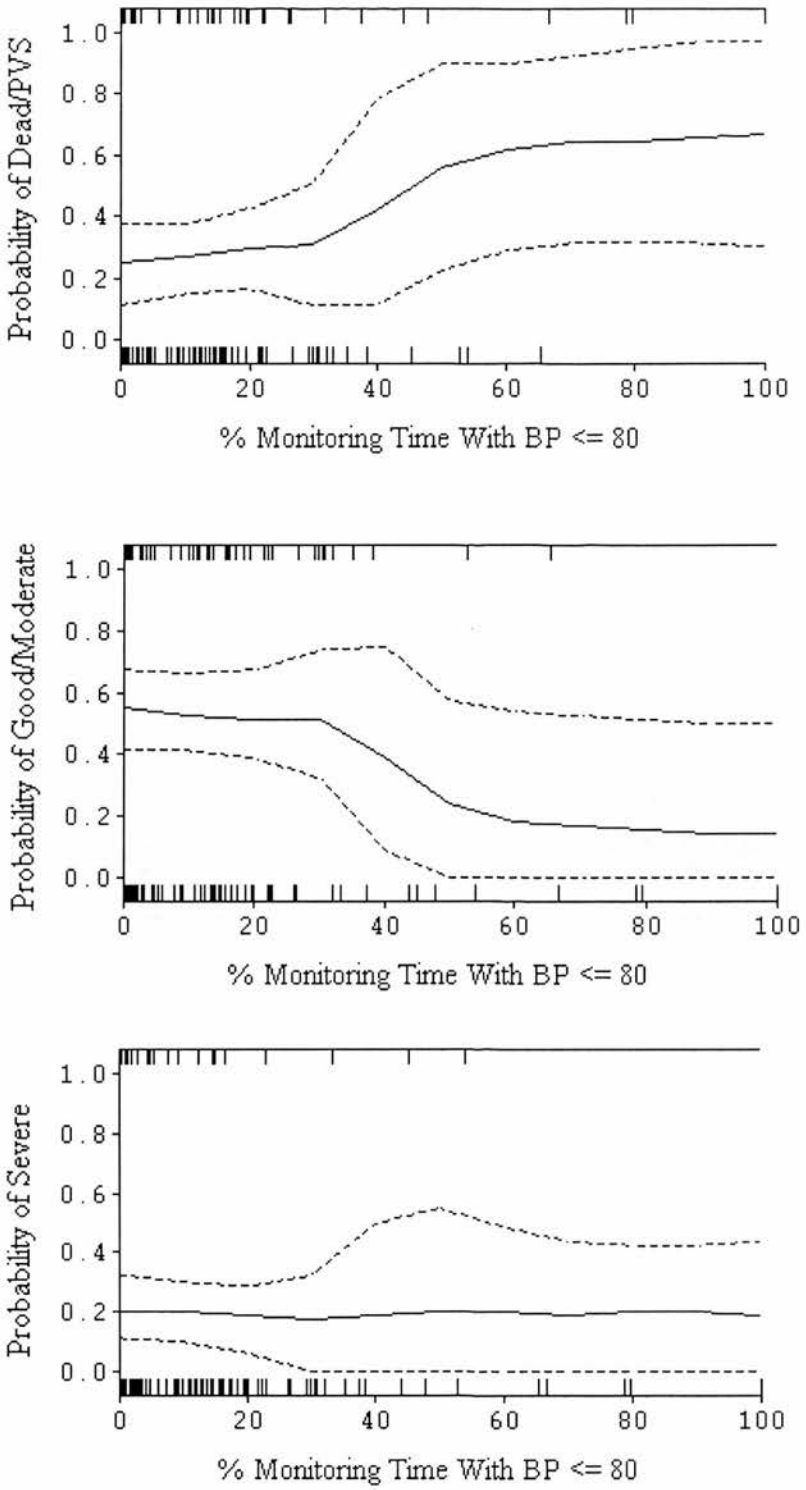


Figure 15 Neural network outcome predictions given proportion of monitoring time spent with arterial blood pressure below 80 mm Hg. The plot for the severe class is almost flat, indicating little or no effect from arterial hypotension.

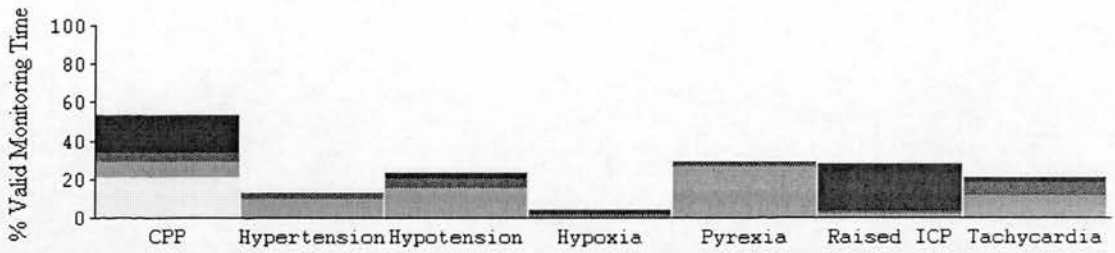


Figure 16 The associated secondary insults for the patient group with arterial blood pressure below 80 mm Hg more than 30% of valid monitoring time (18 patients).

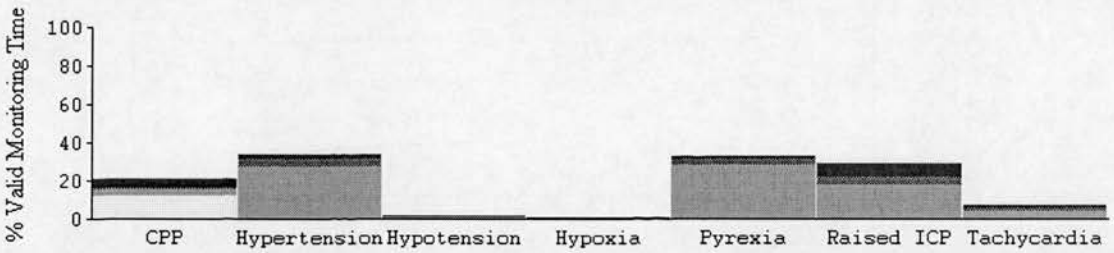


Figure 17 The associated secondary insults for the patient group with arterial blood pressure below 80 mm Hg less than 30% of valid monitoring time (100 patients).

these patients died. If we look at outcome distributions for the three patient groups (figures 18 - 20) we can see that the significance of hypotension as a predictor of outcome in this data set is entirely attributable to these four patients. Since these four were very severely injured and also had large amounts of ICP and CPP insult, there is no evidence here that arterial hypotension makes a contribution to poor patient outcome independent of its contribution to a reduction of CPP.

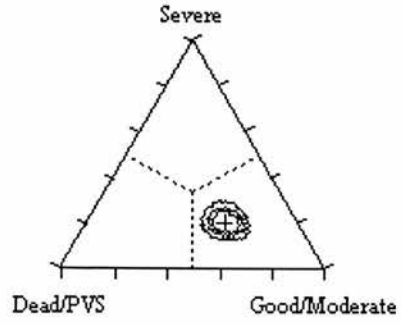


Figure 18 100 patients with blood pressure below 80 less than 30% of the time

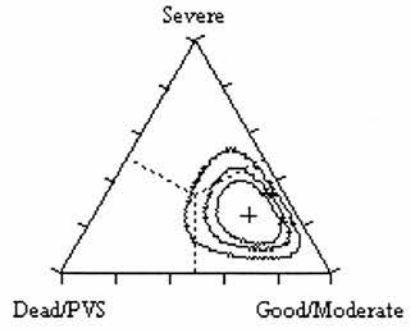


Figure 19 12 patients with blood pressure below 80 more than 30% of the time but at not at high risk from both ICP and CPP

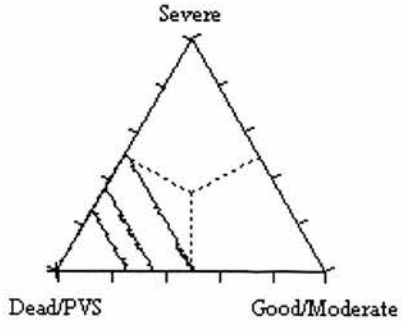
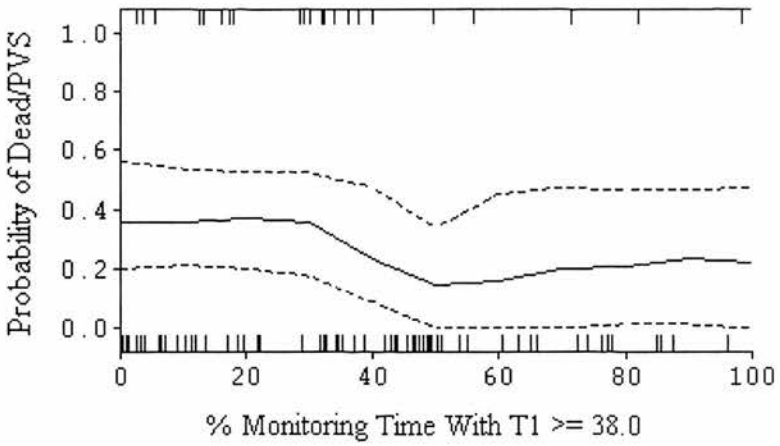


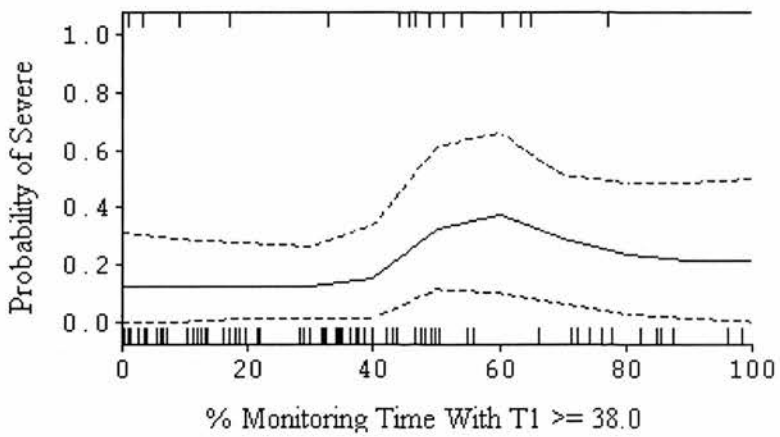
Figure 20 four patients with blood pressure below 80 more than 30% of the time and at high risk from both ICP and CPP.

5.5 Pyrexia

Pyrexia has been shown to be related to survival following head injury (Jones et al, 1994, Signorini, 1999b). Unexpectedly, high levels of pyrexia have consistently been found to be predictive of survival rather than of death. The neural network estimates shown in figure 21 provide some insight into this anomaly. As the proportion of monitoring time spent with core temperature elevated above 38 degrees is moved from zero to one, the probability of good to moderate outcome remains almost constant. However, the probability of death exhibits the step function behaviour we saw with CPP. The step down in probability of death is accompanied by a step up in probability of severe disability. Compared with the probability estimates we have seen previously, these all look rather "jumpy" and indecisive. Still, consistent with previous work, there does appear to be an association of pyrexia with survival, albeit survival with severe disability.

Referring to figure 21, we can see that the point at which the association of pyrexia and survival starts to be in evidence is when temperature is elevated above 38 degrees for about 30% of monitoring time. Breaking our patient groups down using this threshold we find 55 patients in the high temperature group and 48 in the low temperature group. The rates of pyrexic and associated secondary insults for these two groups are plotted in figures 22 and 23. Again, the insults are coded according to insult severity with the more severe being darker. The most salient difference between these two groups is the prevalence of very severe ICP and CPP insults associated with low temperature. In the previous section we saw that the grade 3





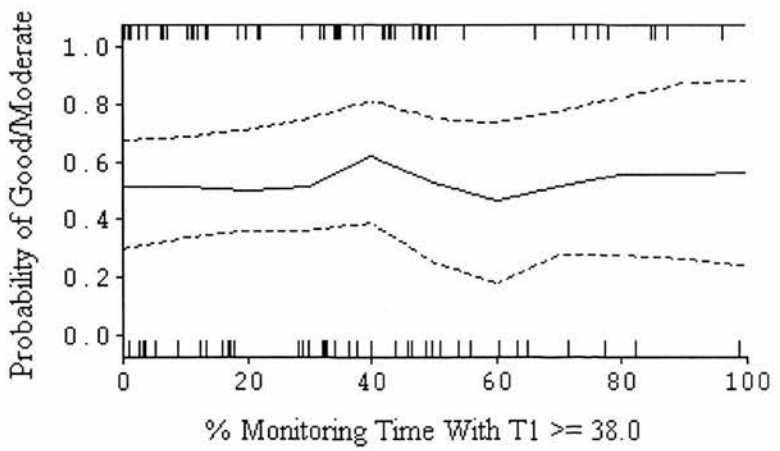


Figure 21 The output estimates of the neural network system given proportion of monitoring time with core temperature over 38 degrees.

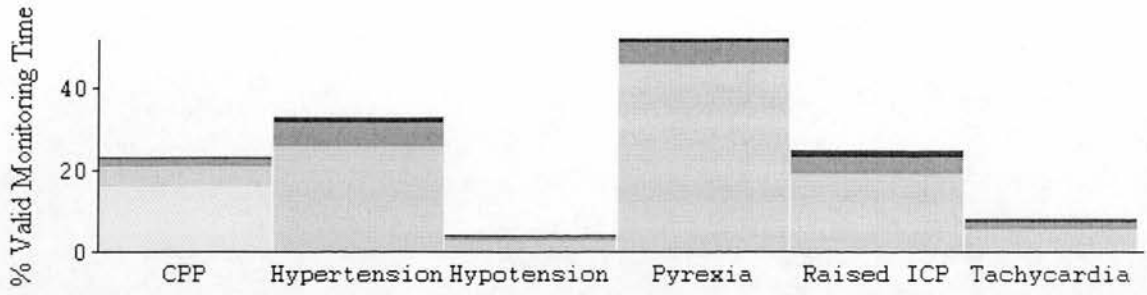


Figure 22 Associated secondary insults for the 55 patients with over 30% of monitoring time with core temperature over 38 degrees.

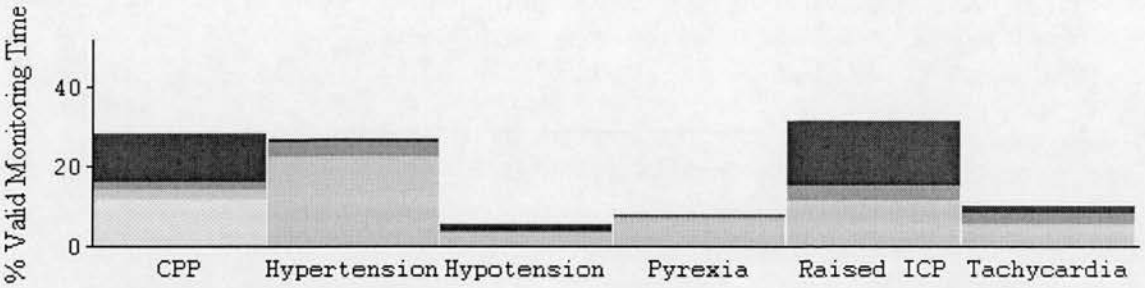


Figure 23 Associated secondary insults for the 48 patients with less than 30% of monitoring time with core temperature over 38 degrees

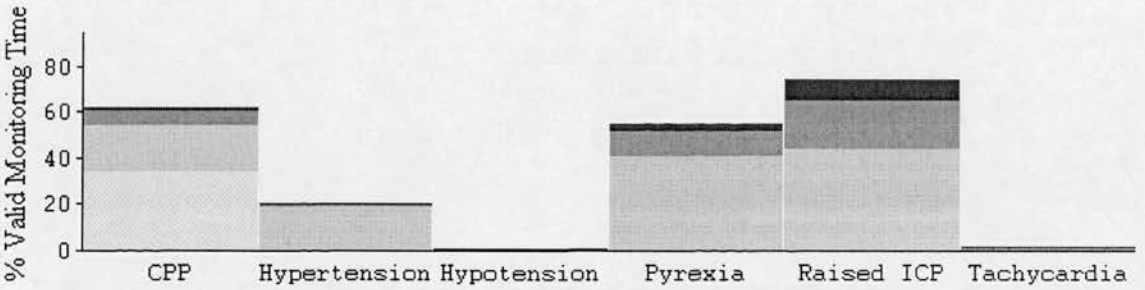


Figure 24 The insult distributions for the high risk ICP/high risk CPP patients who had core temperature over 38 for more than 39% of monitoring time (8 patients)

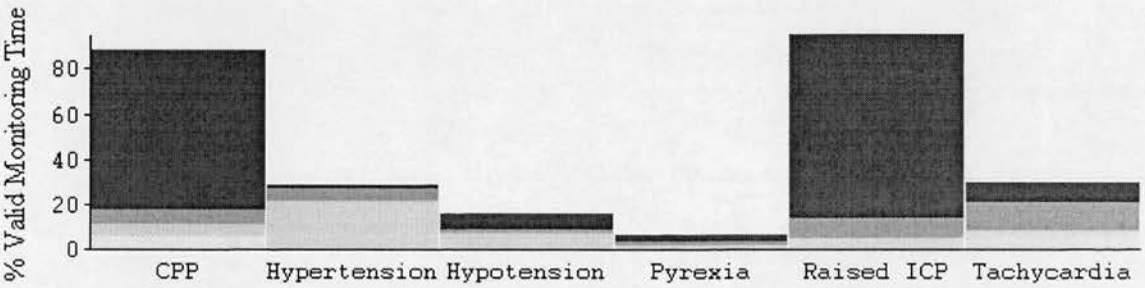


Figure 25 Insult distributions for the high risk ICP/high risk CPP patients who had core temperature over 38 for less than 39% of monitoring time (6 patients)

insults for ICP and CPP are almost entirely concentrated in a group of 17 patients identified as those with ICP over 20 and CPP under 70 more than 40% of the time. If we look at the association of pyrexia with other insult categories confined to this group (figures 24 and 25), the strong association of low amounts of pyrexia and high amounts and severity of other insults is even more obvious.

Pyrexia (or its absence) appears to be much more significant for its association with other syndromes than it is in itself. In patients at high risk from raised ICP and lowered CPP, the absence of pyrexia is strongly associated with the most extreme cases and with impending death. There is little evidence here for an independent contribution to outcome, adverse or otherwise.

5.6 Summary

The analysis applied in this chapter begins with the use of the neural network system as a flexible nonlinear model that relates the duration of secondary insults to patient outcome. This allows us to identify groups of patients who are at risk or are not at risk from particular insults. For example, when we look at the ICP model (figure 3), we see that ICP is not affecting outcome until percentage of monitoring time with ICP over 20 reaches about 40%. We see a similar effect with the occurrence of CPP insults based on CPP falling below a threshold of 70 mm Hg (figure 4). This allows us to define four different patient groups: those with both ICP and CPP problems, those with neither problem, and those with one or the other

In comparing the outcome distributions for these four groups (figure 26) we see that, for the patients studied here, whle ICP in conjunction with a reduction in

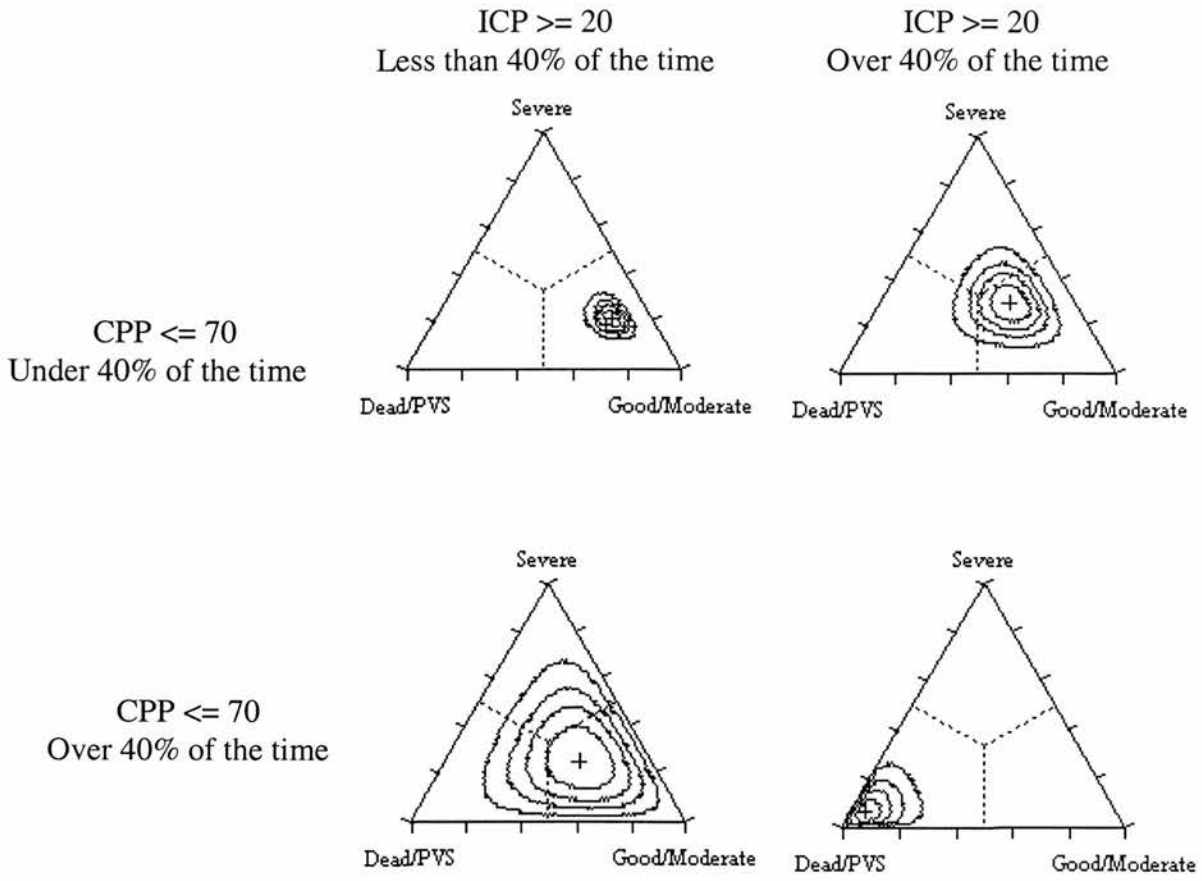


Figure 26 The whole is greater than the sum of its parts. Outcome statistics for four patient groups:

- *Upper left*: 76 patients at low risk from both ICP and CPP
- *Upper right:* 17 patients at high risk from ICP but not CPP
- *Lower left:* 4 patients at high risk from CPP but not ICP
- *Lower right:* 17 patients at high risk from both ICP and CPP

CPP is highly significant, neither factor without the other has a great effect on outcome. In the high risk group (CPP <= 70 more than 40% of the time *and* ICP >= 20 more than 40% of the time), 15 of the 17 patients (88%) died.

Despite being managed according to a strict CPP maintenance protocol, these patients manifested both long durations and extreme severity of CPP, ICP and other insults (figures 9 and 10). This raises the concern that they are untreatable. However, based on their admission data, these patients were not much, if at all, worse off than the other groups (table 2).

Looking at the time course of the occurrence of secondary insults (figure 12), we see that for all patient groups there is an increase in ICP insult over the first four to five days following injury. However, it is only the high risk group which suffers increased amounts of CPP insult. In the first day following injury, these patients are distinguished primarily by a relatively high rate of *very severe* ICP and CPP insult, and the gap in this respect between the high risk patients and the other groups widens rapidly in the early days following injury. On the other hand, the lowest grade of insult considered here for CPP (CPP <= 70 and > 60 mm Hg) appears to have no effect at all on outcome, and its inclusion can even obscure the very real effects of the higher insult grades.

If we look at the third column of figure 12, we see that problems with CPP are either rapidly eliminated or they run out of control. Keeping in mind that this only happens if the patient also is experiencing problems with raised ICP, this suggests the operation of a feedback loop involving ICP and CPP. Rosner (1985) has described exactly such a process underlying the loss of CPP autoregulation, which he calls "the vasodialatory cascade":

> "As CPP is reduced, vasodilation occurs, which is accompanied by an increase in cerebral blood volume. This then leads to an increase in ICP which further reduces the CPP. Unless CPP is restored in some manner, the cycle continues until vasodilation is maximal."

It seems likely that the "high risk ICP/high risk CPP" patient group described in this study were the victims of some kind of self perpetuating process like the one described by Rosner. The extremely high mortality in this group underlines the necessity of halting the cycle at an early stage.

I have included arterial hypotension and pyrexia in this section, because this analysis shows that they are only significant in this database for their association with severity of injury and the most extreme grades of ICP and CPP insult. The *absence* of pyrexia is associated with the occurrence of grade three ICP and CPP insult (figures 22 and 23). There is no apparent contribution of hypotension to outcome beyond its association with four patients who were very severely injured and who also had very severe ICP and CPP insults (figures 18 through 20).

Based on this data, it appears that in managing patients with traumatic brain injury, ICP, arterial hypotension, and pyrexia are best considered in relation to reduced CPP rather than as independent syndromes. The most important clinical factor to monitor is a reduction of CPP below 60 mm Hg in conjunction with raised ICP. It is particularly important to monitor for this condition in the first few days following injury. In the patients studied here, this condition was either virtually completely eliminated by the second day following injury, or it spiralled out of control. Mortality in the latter patient group was 88%.

Chapter 6

Clinical factors relating to cerebral oxygenation

This chapter will continue the analysis of the Edinburgh head-injury database using Bayesian neural networks. The previous chapter concentrated on cerebral hemodynamics, and the significance of cerebral perfusion pressure. This chapter will look at brain oxygen supply and metabolism. This can be approached indirectly by measuring arterial and venous blood oxygen saturation levels.

6.1 Hypoxia

Arterial blood oxygen saturation (SaO2) levels can be measured non-invasively through the use of infrared sensors clipped on a finger or an ear lobe. This technique is widely used in intensive care to help ensure that the patient is adequately oxygenated. The Edinburgh

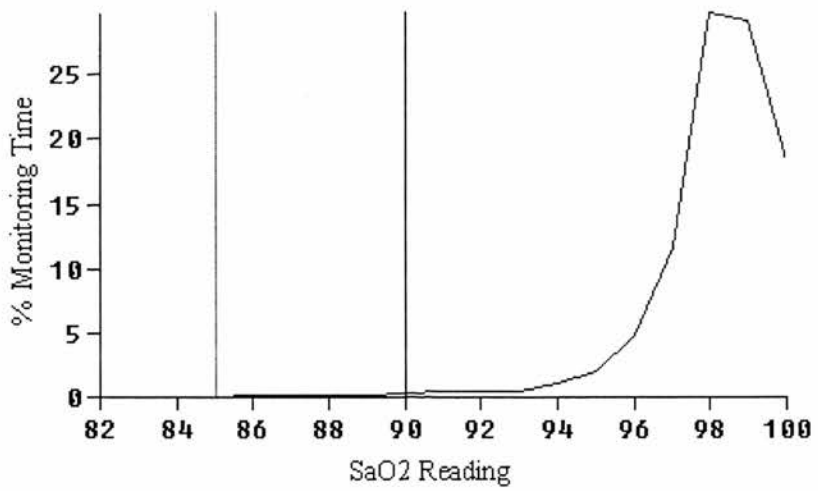


Figure 1 SaO2 Sampling distribution for the Edinburgh database. The horizontal axis is percent SaO2 saturation, and the vertical axis percentage of monitoring time

University Secondary Insult Grades define the grade one insult threshold for SaO2 as 90% or less oxygen saturation. This leads to a problem for analysis of this parameter because there is very little data in this range. Figure 1 is a histogram of the SaO2 sampling distribution for the entire Edinburgh database. The horizontal axis is SaO2 level, and the vertical percentage of monitoring time. This shows how little data is recorded below the 90% threshold. Figure 2 demonstrates how this translates into a distribution for this feature over the patients in the database. This figure is a histogram of the numbers of patients that fall into binned ranges on percentage of monitoring time for this parameter. Almost all of the patients are concentrated on or near the zero percent level with a few scattered elsewhere. This means that using this threshold produces a feature that simply flags a few patients as being special rather than registering gradations of severity smoothly over the patient population. It is difficult to assess the reliability of a feature like this, because it depends so heavily on such a small number of patients. Changing the database just slightly might

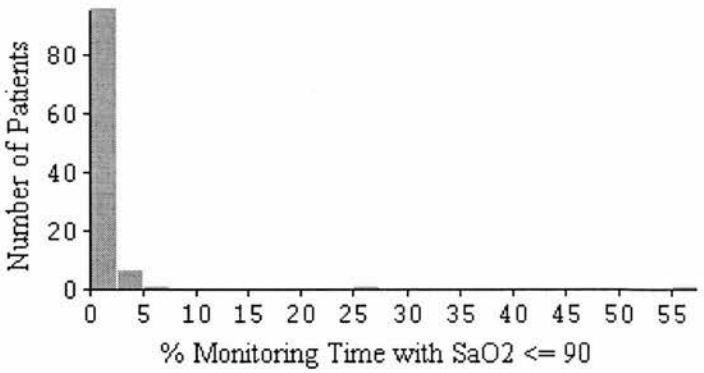


Figure 2 Histogram of the distribution of percentage of monitoring durations for arterial oxygen saturation levels under 90%. The horizontal axis is percentage duration broken into bins representing 2.5% increments. The vertical axis is numbers of patients per bin. Almost all of the patients are in the range 0% to 5% of monitored duration for this parameter

produce completely different statistics. In this instance, there were three cases in which the percentage of monitoring time was greater than 5%, and all of these patients died.

Figure 3 shows the predictions for a neural network system trained with this feature as its only input. The system was given inputs ranging from 0% to 100% of

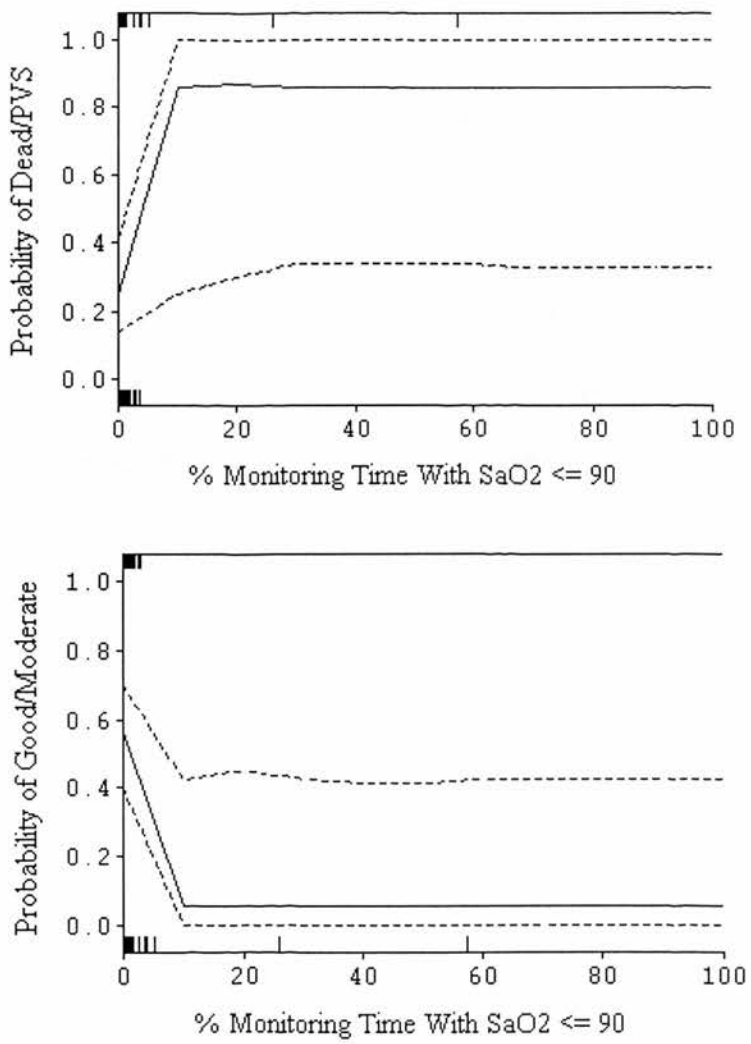


Figure 3 A poor choice of threshold for determining a feature: Outcome probabilities conditioned on the proportion of monitoring time with less than 90% oxygen saturation

monitoring time with SaO2 below 90% (horizontal axis), and the predicted outcome probabilities are graphed against the vertical axis. The neural network system accurately captures the structure of the data, jumping from predicting approximately the prior probabilities to a strong prediction of death. It also puts appropriately broad confidence intervals on these latter predictions which are entirely based on only three patients. Looking back at figure 1, a more appropriate choice of threshold might have been the 96% saturation level. This incorporates enough of the tail of the distribution that we are likely to see a more reasonable distribution over patients than was the case with the 90% threshold. The histogram for the new parameter is shown in figure 4. This has a better spread over the population of patients, and promises to be a more interesting parameter to use for analysis. The predictions for the neural network system are shown in figure 5. Again, the system is making a series of predictions given as input percentage of monitoring time during which oxygen saturation is under threshold. Now we can see more detail in the relationship between reduced oxygen saturation levels and outcome. There is a clear linear trend for increasing

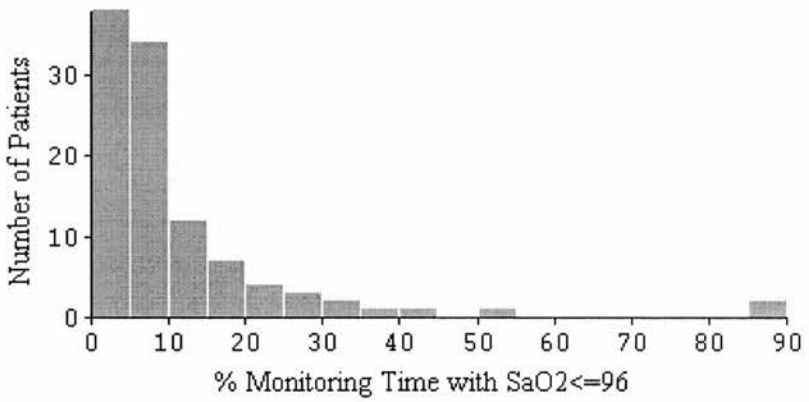
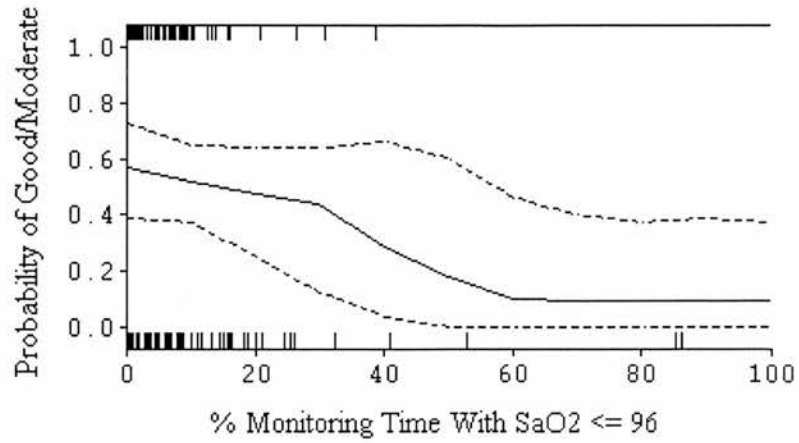
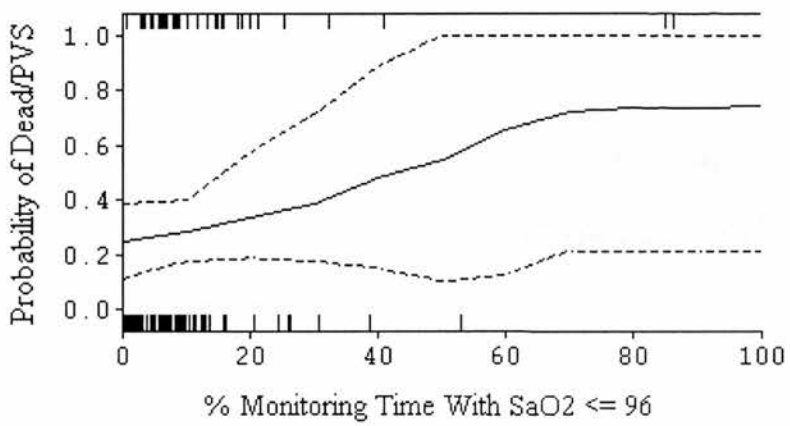


Figure 4 Histogram of the distribution of percentage monitoring durations for arterial oxygen saturation levels under 96%.





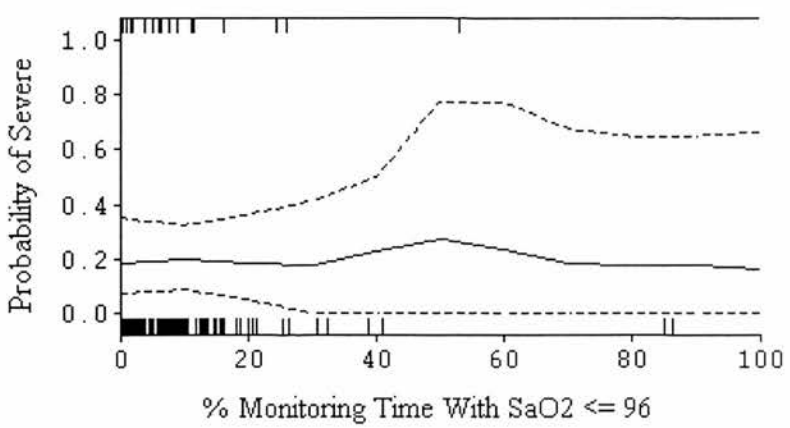


Figure 5 A better choice of threshold for determining a parameter: Outcome probabilities conditioned on the proportion of monitoring time with less than 96% oxygen saturation

risk of mortality and decreasing probability of good to moderate outcome that is evident even when the recorded periods of desaturation are brief. This trend continues in a consistent manner well into the range of prolonged periods of insult. Although the error bars are still wide in the region of long duration insults, most of this uncertainty is between whether the patient will die or suffer severe disability. The feature selection of the 96% saturation threshold was purely based on the distribution of this feature in our data set. However, its utility as a prognostic feature may have implications for clinical practice.

6.2 *Cerebral Hyperemia*

Venous oxygen saturation of blood leaving the brain (SvO2) can be measured by inserting a catheter into the jugular bulb. This can give some insights into the brain's oxygen metabolism. For example, it has been hypothesised that unusually high SvO2

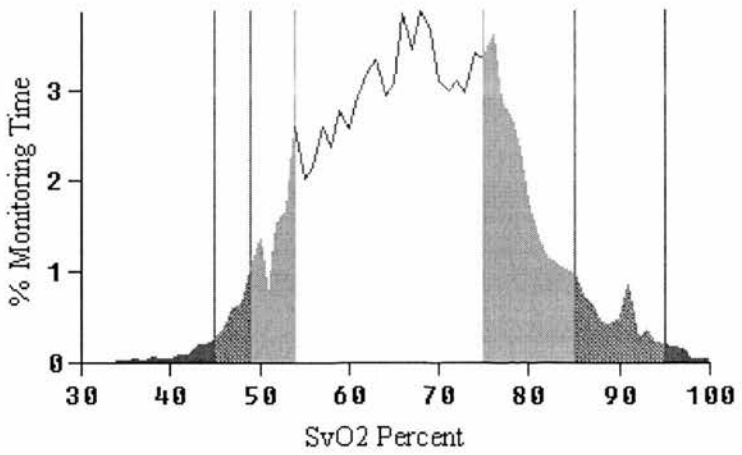


Figure 6 The sampling distribution for jugular bulb oxygen saturation levels. The vertical axis is percentage of monitoring time, and the horizontal percentage SvO2.

levels indicate that the brain's oxygen metabolism is impaired due to neuronal damage. On the other hand, unusually low SvO2 levels have also been linked to brain damage. A high rate of oxygen consumption in the brain may be associated with trauma and attempts to repair nerve damage. Figure 6 shows the sampling distribution for SvO2 levels in the Edinburgh database. We will be looking at levels below 54% and above 75%, which are the EUSIG grade one insult thresholds.

Unfortunately, using the criteria used for the other parameters we have looked at for inclusion, i.e. at least 6 hours of valid monitoring time within the first 48 hours following injury, we only get 42 patients to study for this parameter. This is mainly due to technical problems with the sensors that often made it impossible to ensure that the data was valid. It is difficult to place the sensor correctly in the jugular bulb, and to validate that it is correctly placed. The sensor positioning is also sensitive to any movement of the patient. Much of the data we collected had to be discarded for this reason.

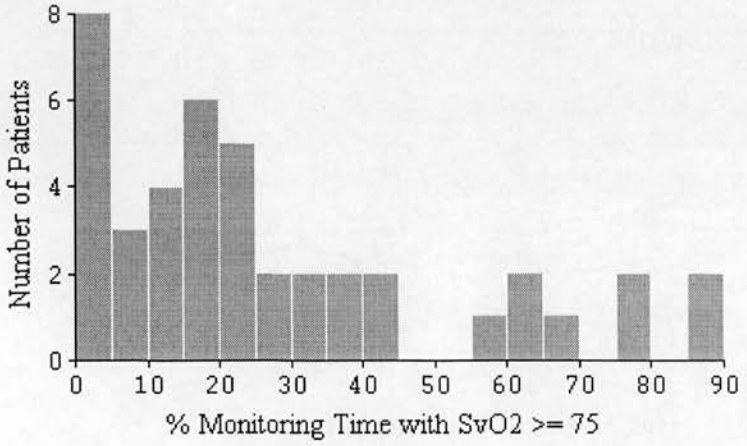
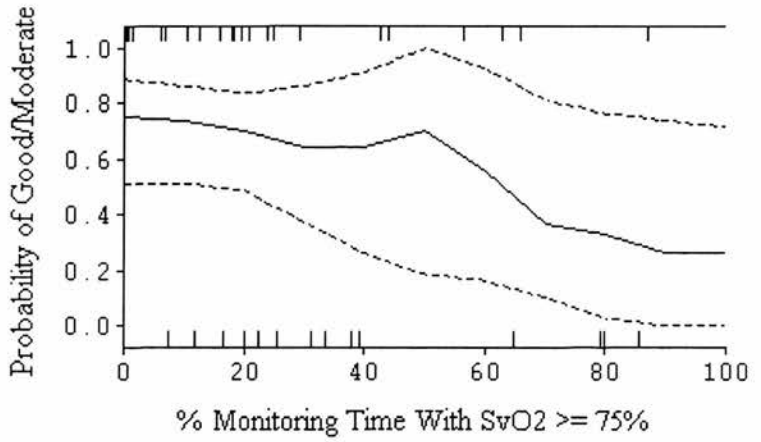
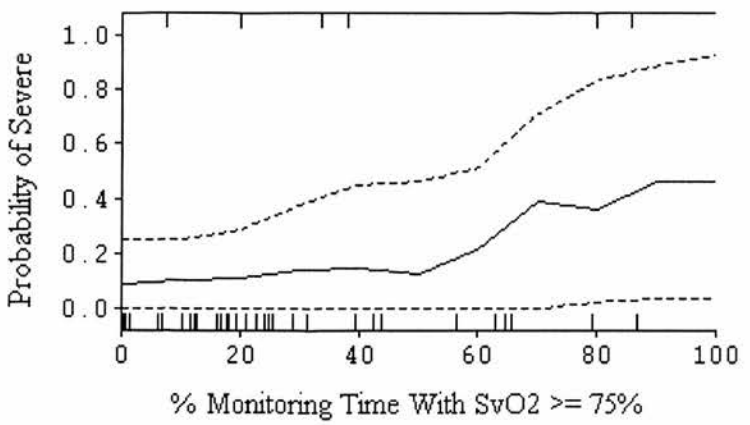


Figure 7 The distribution of the parameter representing percentage of monitoring time that SvO2 is over 75%. The vertical axis is numbers of patients and the horizontal percentage of monitoring time binned in increments of 5%.





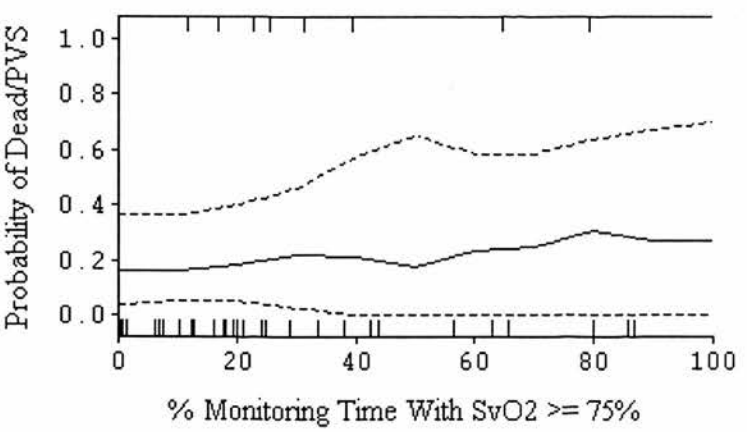


Figure 8 There appears to be an association between high SvO2 levels and poor outcome, despite the small sample size.

The feature used here to represent the occurrence of high SvO2 levels, or cerebral hyperemia, is percentage of monitoring time for which SvO2 is over 75%. The distribution of this parameter over the patients in the Edinburgh database is shown in figure 7. Figure 8 shows the output of the neural network system trained on this feature. The error bars are very broad due to sparse data. However, there does seem to be a clear trend of increased risk for severe disability or death with increasing duration of periods of hyperemia. As was the case with arterial oxygen desaturation, this trend appears to be roughly linear, and is evident even for relatively short insult durations

6.3 Cerebral Oligemia

Abnormally low jugular bulb levels are believed to indicate that the brain is extracting high levels of oxygen in an attempt to recover from trauma. This condition has previously been shown to be associated with poor outcome following head injury (Gopinath, et al., 1994). However, I did not find any interesting models relating cerebral oligemia to outcome, possibly because of the small sample size available (42 patients).

6.4 The time course of SvO2 insults

When the occurrence of periods of high and low levels of jugular bulb oxygen saturation are plotted on a daily basis, a clear pattern is evident. The duration of oligemic insults steadily decreases in the days following injury, while the duration of hyperemic episodes increases (figure 9). These results are for 27 patients who each had at least one hour of valid SvO2 monitoring for each of the first 24 hour periods

following injury. These results are consistent with the theory that in the first stages following brain trauma, the brain's oxygen metabolism rate is unusually high due to attempts at self repair. Then it becomes unusually low due to cell death. Similar results reported are in (Cormio et al, 1999)[1]. This study found that high SvO2 levels can have diverse causes. However, the dominant cause was a decrease in the brain's oxygen metabolism. This study also found an association between elevated SvO2 levels and poor outcome, as did another study by Macmillan and colleagues (1998).

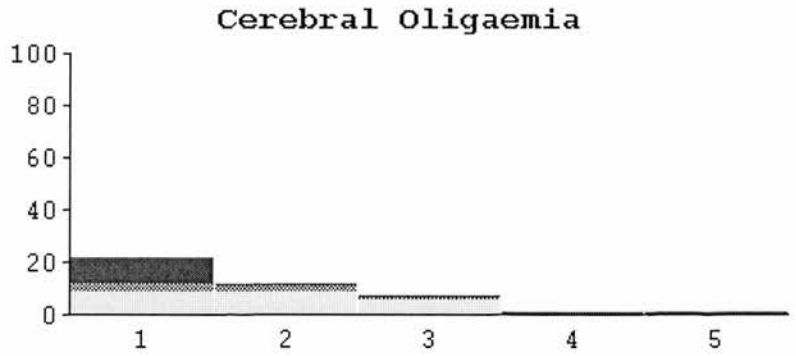


Figure 9 The time course of oligemic insults. Vertical axis is percent of monitoring time. Horizontal axis is day following injury. The insult thresholds used are 54%, 49%, and 45% oxygen saturation (SvO2).

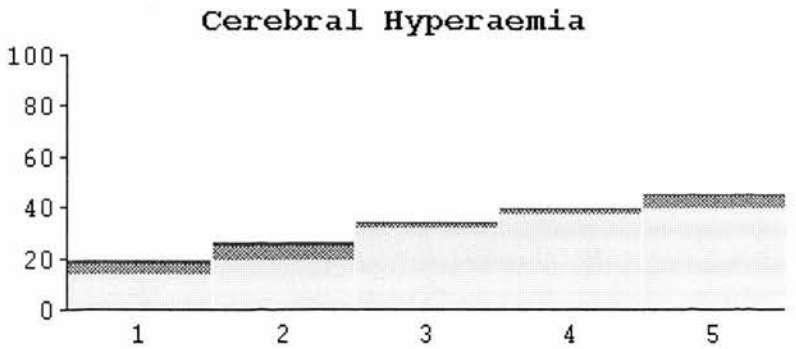


Figure 10 The time course of hyperemic insults. Vertical axis is percent of monitoring time, horizontal axis is day following injury. The insult thresholds used are 75%, 85%, and 95% oxygen saturation (SvO2).

[1] Thanks to Giuseppe Citerio for pointing this work out to me and suggesting the time oriented analysis.

6.5 Summary

In this chapter we looked at two parameters relating to the supply and metabolism of oxygen in the brain. Arterial oxygen saturation (SaO2) is of particular interest clinically because it is a very easy parameter to monitor, understand, and control. The initial experiment based on SaO2 values under 90%, however, was primarily of interest as a test of the neural network. The distribution of this parameter over patients was very peculiar and posed a challenge to the modelling technique that was passed with flying colours. The threshold of 90% had been selected because of its use in defining the lowest category of hypoxic insult in the Edinburgh University Secondary Insult Grades. To obtain a better distribution over patients, a second model was trained using a threshold of 96%. This threshold would generally be considered quite high from a clinical point of view. However, the model shows a clear linear decrease in probability of good outcome as the duration of monitoring time under threshold increases. This is evident even for short time periods under threshold.

In this chapter we also looked at jugular bulb oxygen saturation (SvO2), which tells us how much oxygen remains in the blood as it leaves the brain. This parameter is difficult to monitor, and problems with one of the types of monitors we used lead to the loss of much of the data we collected. Despite a small sample size it was possible to demonstrate an association of high SvO2 levels with poor outcome, which is consistent with other recent work. The time course of SvO2 insult is interesting, with very low levels dominating at first, which then over a period of about five days give way to very high levels.

Chapter 7

Multivariate models combining physiological monitoring data and admission data

In chapter three I compared a number of multivariate Bayesian neural network models. All of these were based on demographic data and simple clinical indicators available when the patients were admitted to neurosurgery. The models discussed in chapters five and six were based on the physiological monitoring data, but these were univariate. In this chapter we will look at the performance of multivariate models that incorporate physiological data together with the prognostic factors that proved to be most important in chapter three. This will allow us to examine the performance of these models with higher dimensional input spaces and sparser data. It will also lend support to some of the hypotheses developed in the previous two chapters to the extent that the physiological factors identified as being important lead to models that perform better than models based on admission data alone, or on other physiological factors.

7.1 Model Comparison

The thresholds used to determine the physiological parameters in the models I tried are shown in table 1. The models tried are listed in table 2. I have only included patients over the age of 14 who are classified as having severe head injuries. All of the models include age, pupil score, and motor score, which were found in chapter three, as in much previous work, to be the most important indicators available on admission. I have included patients in the training set for a model if all of these are

Table 1 The thresholds used for calculating proportion monitoring time over or under threshold for the physiological channels used in these models

PHYSIOLOGICAL CHANNEL		CONDITION
ICP	Intracranial Pressure	>= 20 mm Hg
BP	Arterial Blood Pressure	<= 80 mm Hg
CPP	Cerebral Perfusion Pressure	<= 60 mm Hg
Sa02	Arterial Oxygen Saturation	<= 96%

Table 2 The multivariate models tried, listed in order by logarithmic error, from best to worst. All models included age, motor score, and pupils score. The thresholds used on the physiological channels are given in table 1. N is training set size. The delta errors are the improvements over guessing the most common class achieved by the model. All errors are based on 10-way cross validation. The errors are based on the different training sets available, and are therefore not directly comparable.

MODEL	N	PERCENT ERROR	PERCENT ERROR DELTA	LOGARITHMIC ERROR	LOGARITHMIC ERROR DELTA
ICP CPP	103	0.431	+0.025	0.904	+0.106
CPP	103	0.422	+0.035	0.910	+0.099
ICP BP	103	0.437	+0.025	0.912	+0.100
ICP	103	0.398	+0.064	0.915	+0.098
Big Demographic	242	0.402	+0.100	0.965	+0.047
CPP SaO2	87	0.425	+0.036	0.969	+0.036
Small Demographic	103	0.341	+0.069	0.978	+0.031

available, and the patient had at least 6 hours of valid monitoring time within the first 48 hours following injury on each of the physiological channels included in the model. The input features for each physiological channel were calculated as the proportion of monitoring time with that channel above or below a given threshold. The thresholds used are listed in table 1. These are based on previous work as discussed in chapter four, together with the modifications suggested in chapters five and six. In addition to the models incorporating physiological data, I have tested two models that only use age, motor score, and pupil score. One of these, "Big Demographic", is the one described in chapter three, which was trained on the full

Table 3 Performance of the pupils/motor/age model trained on the full data set (N=242) evaluated on the full data set (first row), and the data set with ICP and CPP data available (N=103, second row). Apparently the patients being intensively monitored are less predictable.

TEST SET	N	PERCENT ERROR	LOGARITHMIC ERROR
Admission Data Only	242	0.364	0.832
With Physiological Data	103	0.402	0.965

data set (242 patients) available with this data. The second, "Small Demographic", was trained on the 103 patients available for most of the other models. These patients all had sufficient ICP, CPP and BP data to be included in the training sets for models using those parameters. Only one model had fewer patients. This was the model based on CPP and SaO2, for which only 87 patients qualified

One interesting result is that the performance of the "Big Demographic" model is worse on the subset of patients with physiological data than on the full data set, as summarised in table 3. Apparently the subset of patients being intensively monitored are the most unpredictable. This is as it should be. ICP monitoring is invasive, and carries the risk of infection. Apparently it is being used appropriately to monitor the patients who are most unstable, and therefore most likely to benefit from intensive monitoring. This result is consistent with the findings of a previous statistical analysis of this database (D. Signorini, personal communication).

The sample sizes available for training these models are much smaller than those used to train the simpler models in chapter three. Even given the larger data sets available for the simpler models, it was evident that percent error was an unreliable metric due to the numbers of predictions near decision boundaries. In this case, I would not put any weight on the percent error, although I have reported it in

table 2. Based on the more reliable logarithmic score, I might claim some degree of confirmation of the results reported in chapter five. There my analysis indicated that the combination of ICP and CPP insult is the most critical factor for these patients. Here we find that the logarithmic score for the model incorporating ICP and CPP is the best of all models considered, including the "Big Demographic" model, which had a training set more than twice the size of the one available to the models utilising physiological data. In fact *all* of the models that incorporate physiological data outperform "Big Demographic" with the sole exception of the CPP/SaO2 model. This model had the smallest training set of all (87 patients).

However, it is by no means clear that the differences in logarithmic score between any of these models are significant given the very small sample sizes. In fact, a closer look indicates that like percent error, the logarithmic scores may be misleading in this case. A possible explanation for the poor performance of "Big Demographic" is evident from an inspection of model predictions on a case by case basis. This model predicts more confidently than the models trained on smaller data sets. Given this relatively unpredictable group of patients, there are cases in which it predicts the wrong class with high certainty. Logarithmic error diverges to infinity when the probability assigned to the correct class goes to zero. Therefore a few overconfident predictions can disproportionately damage a model's score.

A graphical display for inspecting the separation of outcome classes achieved by models was introduced in chapter three. This plots the mean prediction made by the model for patients in each of the three outcome classes. The plots for the seven models discussed here can be seen in figure 1. Judged in this way the "Big Demographic" model appears to be the best.

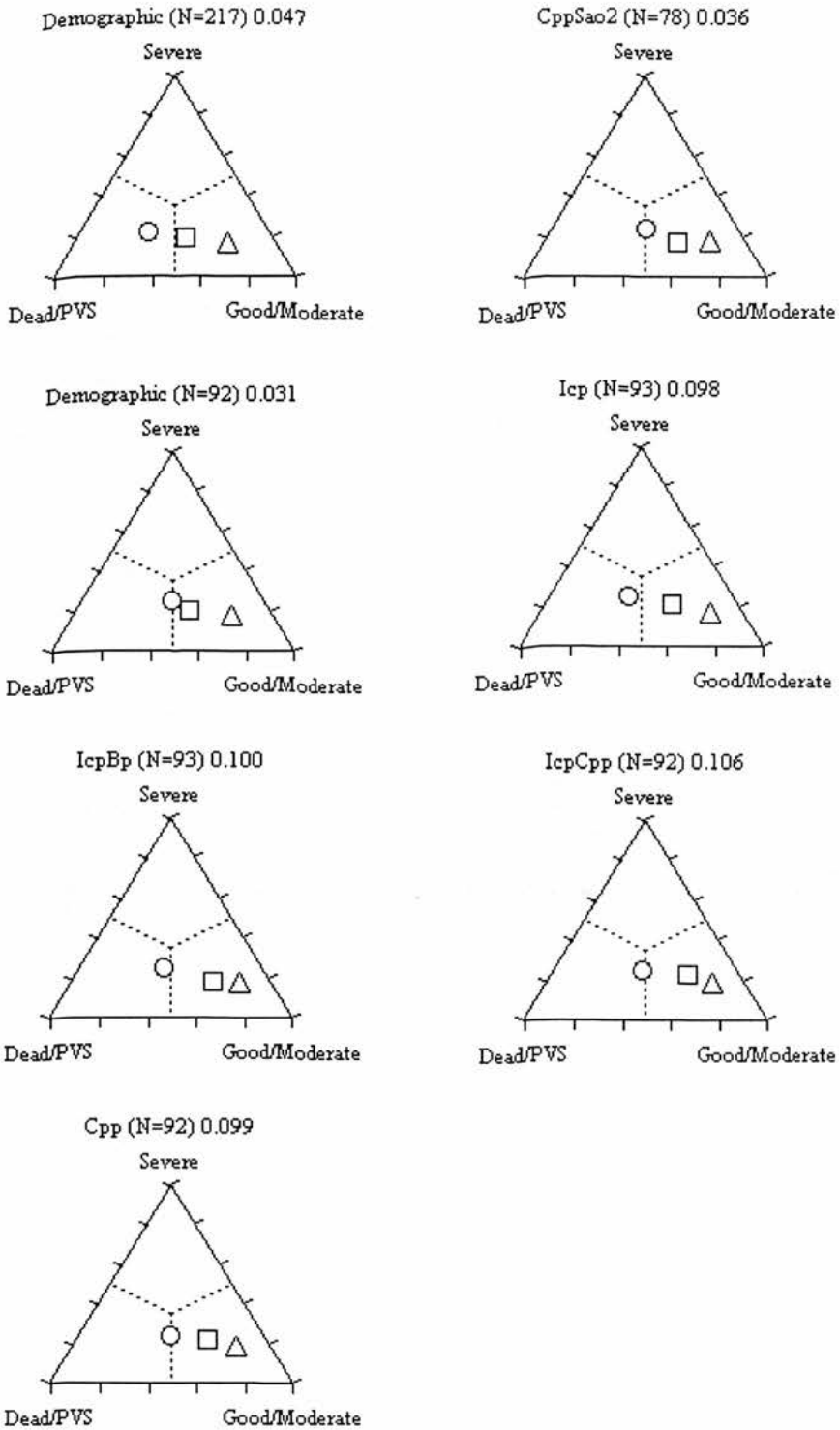


Figure 1 Model separation of outcome classes for the seven models. These plot the mean vectors predicted for the three outcome groups: Death/PVS (circle), Severe Disability (square), and Good Outcome/Moderate Disability (triangle). The "N" values here are the sizes of the cross validation training sets - not the full sample size.

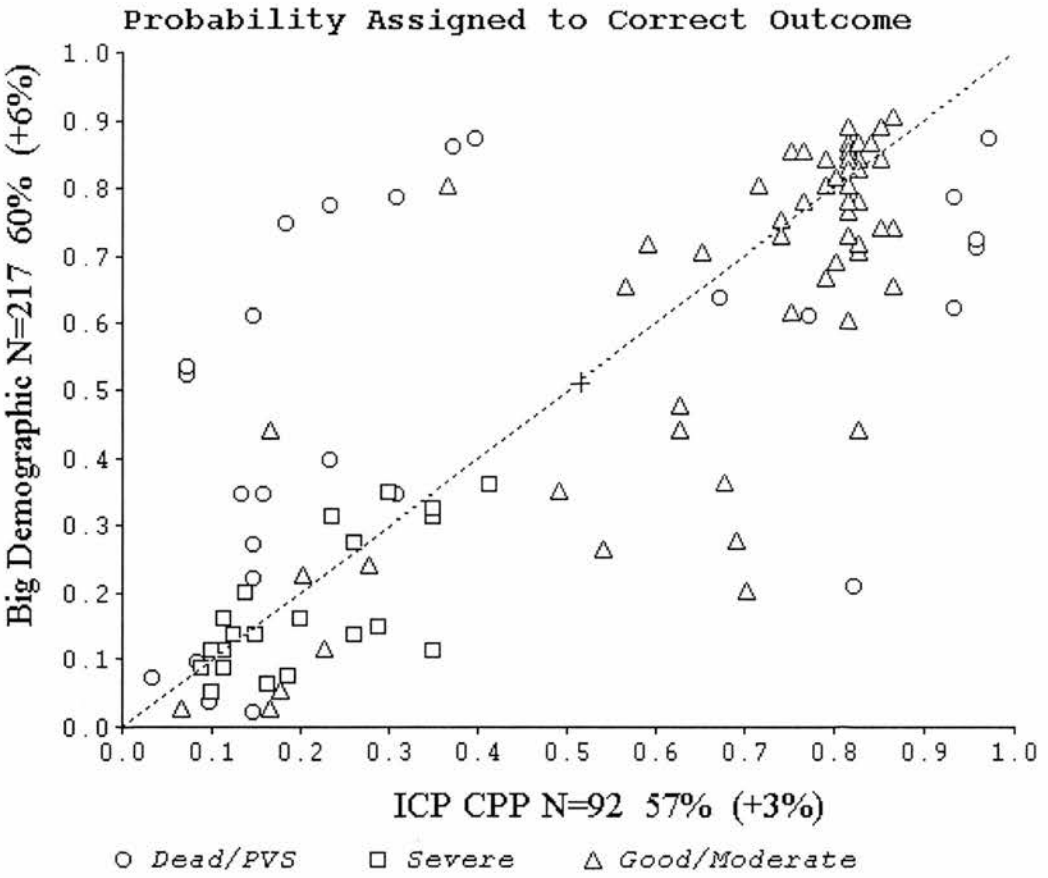


Figure 2 The predictions of "Big Demographic" compared with those of the ICP/CPP model. Big Demographic does better on a cluster of eight patients who died (upper left).

To compare model predictions in more detail we can make a scatter plot of the probability assigned to the correct class by two different models over the test cases. Figure 2 plots this for the ICP CPP model and "Big Demographic". The general trend of the predictions favours ICP CPP with the very notable exception of a cluster of eight patients who died (they're in the upper left of the plot). These cases were much better classified by Big Demographic. These patients were very severely injured. Six had bilaterally fixed pupils, and the remaining two had one fixed pupil. However, all of these patients had relatively low amounts of ICP and CPP insult, which apparently caused the ICP/CPP model to misclassify them. This would seem

to indicate that the model is not learning important exceptions to general rules. This suggests that either there is simply insufficient data in the training sample, or that the model is over-regularised.

7.2 Summary

The results reported in table 2 could be argued to support the results of the previous two chapters, which were based on univariate models. However, the improvements in model performance with the addition of the physiological data over the performance of a simpler model trained on a larger data set are modest and possibly not significant. This is due to two factors that come into play as more inputs are added. First of all, training sets shrink as patients have to be dropped out due to missing data. Secondly, the curse of dimensionality caused by the larger input space will impede learning.

This does not mean that the thesis articulated in chapter two, that Bayesian neural networks can be used effectively in scientific research, was wrong. In chapters five and six Bayesian neural networks were applied effectively in an analysis of this data set through the use of univariate models. It is also encouraging that given this very sparse and complex data set, the multivariate models perform as well as or better than a simpler model trained on a much larger data set. If nothing else, this should persuade us that we are on the right track and inspire us to collect more data.

Two things (at least) can be done to improve the performance of the more complex models. First of all, we *must* get more data! Ultimately there is no substitute for this. Fortunately this is already happening. As described in the

introduction, this study was based on a large scale data acquisition project within a single intensive care unit. The data acquisition software that was developed to support this project is now being used in ten intensive care units across Europe. Soon this will allow the assembly of much larger databases incorporating detailed physiological monitoring and treatment information.

Secondly, I suspect that we have to look carefully at the priors for these networks. This proved to be critical in the development of simpler models that perform correctly given sparse data. It may be that the priors I have used over-regularise when applied to these more complex models, causing them to miss significant features of the data. An aspect of the prior that I have not investigated carefully is its effects on the modelling of interactions between the input variables. This may be an important area for future research.

Chapter 8

Conclusions and Future Work

8.1 Validating the use of Bayesian Neural networks in scientific research

In the introduction I discussed the application of Bayesian neural networks in support of scientific research. In recent years many claims have been made about Bayesian neural nets which, if true, would make them a powerful data analysis tool in this context. For example:

- They accurately model the probability of the target data conditioned on the input data.
- Model complexity is automatically adapted based on the properties of the training data.
- Confidence intervals are assigned that take into account model uncertainty and sample size.
- The form of the model, i.e. the number of layers of hidden nodes and the number of nodes in each layer, can be selected based entirely on prior knowledge without reference to the data. There is no need for a validation set to determine the numbers of hidden nodes; model complexity is adapted to the training set through the use of hyperparameters that control the values of the smoothing parameters.

For the past several years I have been involved in a project analysing a large, complex data set of physiological time series data collected during intensive care of head-injured patients. The application of Bayesian neural networks to this task has required a painstaking process of model validation, starting with the simplest possible models and building up to realistic models that have provided new insights into the risk factors for these patients.

The implementation of Bayesian neural networks used in this thesis is based on Radford Neal's software (Neal, 1996). This system samples from the posterior

distribution for the neural network weight vector conditioned on the training data using Monte Carlo Markov chain (MCMC) techniques. The results reported in chapter two and three using this implementation have provided detailed confirmation of the claims put forward regarding Bayesian neural networks. Model validation is based on a series of experiments using first simulated and then real data. The simulated problems were kept very simple so that exact solutions for the full posterior distributions in output space could be derived for comparison with the neural network estimates. Kernel density techniques were developed to obtain continuous probability density functions from the discrete densities generated by MCMC sampling, enabling a direct comparison of the desired PDF's. The quality of the approximations produced by the neural network was more than adequate for this application. It was demonstrated that these results relied on the use of an approximately uniform prior on the output function, and a procedure for finding such a prior was discussed.

A second series of experiments validated the smoothing and generalisation properties of the system by modelling the relationship between a few simple clinical variables recorded in the Edinburgh head-injury database and patient outcome. Conveniently, these injury severity scales are designed to be monotonic and roughly linear. Because these variables are coarse grained (three to five categories), it was possible to compare the estimates of conditional probabilities generated by the neural network with those obtained simply by treating each possible value that the input can assume separately as a predictor of outcome based on the statistics of the Edinburgh database. By inspection, the behaviour of the neural network system appears to be correct, although even for these still very simple problems, it is no longer possible to

derive exact solutions for comparison. Finally, a series of multivariate models were tested by training them on standard demographic and clinical indicators that have been the subject of statistical prognostic modelling for many years. These experiments replicated well established results in the field, providing another measure of model validation for this implementation of Bayesian neural networks, this time on realistic problems.

The final "proof" that Bayesian neural networks can be used effectively in a scientific project lies in their actual application to the analysis of the Edinburgh head-injury database. Significant clinical results are reported in chapters five and six that were a direct result of the use of these models as tools for exploratory data analysis. These are new observations based on the detailed behaviour of these flexible, non-linear models.

Based on my experience with this implementation over the past few years as reported in this thesis, the claims for Bayesian neural networks listed at the beginning of this section are correct. The rigorous theoretical foundations of this approach together with practical experience will allow it to take its place as a legitimate tool of scientific research and data analysis.

8.2 Clinical conclusions

The specific clinical conclusions of this study can be found in the summaries for chapters five and six. A few cautions should be noted regarding methodology. This is an observational study. This means that there was no serious attempt to rigorously frame this as an experiment to test hypotheses. Rather, I have treated this an exercise in exploratory data analysis. This approach is a useful way of

generating new hypotheses, or providing support for existing hypotheses given a large and poorly understood database. The evidence produced in this way is not as convincing as might be produced if an appropriate randomised controlled trial (RCT) could be designed to evaluate specific hypotheses. However, this is not always appropriate, or even possible or in the context of clinical care (Black, 1999). I hope that the large scale data acquisition and data analysis methodologies demonstrated in this thesis will play a role in providing sound evidence for clinical practice in cases where the experimental paradigm is not a practical alternative.

8.3 Unresolved issues and a criticism

Two important technical issues regarding the application of Bayesian neural networks in the context of a scientific project remain unresolved. The first is the problem of diagnosing convergence of the Monte Carlo Markov chain to the posterior weight distribution. As described in the introduction, I am convinced that in practice this has not been a problem with the work described here. Several times during course of this study I wondered if the system had actually converged. In some cases I experimented with the parameters of the simulation, and in others I let simulations run for a few days to see if they would find a better solution. In no case did lack of rapid convergence to the posterior prove to be a problem in my experience with this data set. Nevertheless, I would be happier if I could have provided better evidence of convergence. Unfortunately, a brief survey of the literature suggests that I might not have succeeded in this even if it had been one of the major aims of this thesis!

A second problem area is discussed in the summary of chapter seven. This has to do with the evaluation of the priors used for the most complex multivariate models tested on this data set. There is some evidence that these priors may result in the models being over-regularised, and that this may cause them to miss important patterns in the data. This would be an interesting area for future research.

This also leads to a criticism of the practical results obtained in chapter five and six. The role of Bayesian neural networks in producing these results was entirely restricted to the use of univariate models. Although the non-linear properties of these models were critical in producing these results, in hindsight any number of less computationally demanding techniques could have been used. This observation could lead (and has led) to the observation that I have used a sledgehammer to kill a fly. My first response to this would be that at least this is a testimony to the exceptionally high quality of the sledgehammer being used. In the introduction I have discussed the fact that in this thesis I have not been interested in demonstrating the full power of neural networks as non-linear classifiers, which has been amply demonstrated many times over. Rather I am interested in the mathematical properties of Bayesian neural networks, and their performance given sparse data. Secondly, I would note that the fact that the simple models would be the most informative was *only* obvious in hindsight. I fully expected that the most interesting experiments would be those involving multivariate models, and in fact the only reason I initially experimented with univariate models of the physiological data was to calibrate the thresholds and the normalisation procedures being used. It was only when I

was confronted with certain "strange" features of these models that interesting results started to emerge. In a real life application you can never be sure in advance exactly what the relevant model is or how it should be applied. Bayesian neural networks will find simple models of simple data and complex models of complex data. They will accurately characterise the uncertainties in these models. You can't ask for more than that.

8.4 Future Work

This work was based in part on software for automatic data acquisition in the intensive care unit. A few years ago the first version of this system was installed in Edinburgh. Now the system is being actively used in ten intensive care units across Europe. I hope that in the next few years this will lead to the formation of much larger databases than the one available to me in this study. I hope that I will be able to make this data available for research in automatic pattern recognition from several different practical and technical perspectives. It would be hard to imagine a more useful or rewarding application for these technologies.

Acknowledgements

Thanks to my advisors, Peter Ross and David Willshaw, for helping guide me through an interdisciplinary thesis project, and for being patient and impatient at the appropriate times. Also to Peter Ross and Chris Williams for many helpful comments on the final drafts of the thesis that made it a much better document than it otherwise would have been. Thanks to professor Mike Titterington for reviewing drafts of the thesis from a statistical perspective, although all remaining errors are my own, of course. Thanks are also due to my clinical collaborators on the head-injury project: Peter Andrews, Pat Jones, and the late professor Douglas Miller, all of whom contributed to making it a very exciting and inspiring project to work on; and special thanks to Ian Piper for his constant enthusiasm and support for this work from the moment I interviewed for a position on the project to the present moment. Special thanks also to Radford Neal, for making a very impressive software package freely available, and for patiently helping me familiarise myself with it on both a practical and theoretical level. It goes without saying that professor Neal's contributions go well beyond the scope of any particular software implementation. The people who have contributed interesting comments and observations on this work are too numerous to mention, but I will try. For useful technical advice: David Barber, Rich Caruana, David MacKay, and David Spiegelhalter. From a clinical perspective: Peter Alston, Giuseppe Citerio, Per Enblad, Carol MacMillan, Sheena Millar, Mike Robson, Bertil Rydenhag, Elisabeth Ronne-Engstrom, Mike Souter, and Graham Teasdale.

Appendix

A.1 Parzen Density Estimation

The Monte Carlo approach to Neural Network learning described in section 1.4 produces a discrete sample of networks from the posterior, which in turn can be used to generate a sample of outputs for some given input. In this thesis I have frequently wanted to compute a continuous density based on this discrete sample of outputs (e.g. see figure 6, chapter 2). I had to solve this for two class and also for three class problems. In both cases I have done this using Parzen density functions. The density at any given point is computed as the sum of circularly symmetrical Gaussians centred on each point in the sample. The key to making this work is determining an appropriate scaling factor (σ) that controls the width of the Gaussians. If σ is too large, the Gaussians will be very wide, resulting in a function that is too smooth. If σ is too small, the Gaussians will be very narrow, and the function computed will be too rough. For both the two output and three output cases, I have used heuristic formulas for computing σ. I can make absolutely no claims for these formulas beyond the fact that they served my purposes well.

A.1.1 Two Output Case

I have computed the smoothing constant (σ) as:

$$\sigma = 1.35\overline{D}\sqrt{N}$$

Where $\overline{D}$ is the mean distance between neighboring points in the sample and N is the number of points in the sample.

Then the Parzen density function for a point x is defined as :

$$F(x) = \sum_{i=1}^{N} \frac{\exp(-((x-\mathbf{o}_i)^2 / 2\sigma^2))}{\sigma\sqrt{2\pi}}$$

Where N is the number of neural networks generated by the Monte Carlo simulation, $\mathbf{o}_i$ is the output of the i[th] network, and σ is the smoothing constant.

Then I define a grid that subdivides the interval (0, 1) into 200 equal segments and P(x) is computed at each point on the grid. The height of the curve is then scaled so that the area under the curve is equal to one.

A.1.2 Three Output Case

As explained in section 2.11, probability distributions over three outputs are confined to lie on the triangle defined in 3-space by the points (1 0 0) (0 1 0) and (0 0 1). For each point in this triangle I have computed the assigned probability using a Parzen density function. First the smoothing constant (σ) is computed as:

$\sigma = 0.35\overline{D}$

Where $\overline{D}$ is the mean distance between *all* pairs of points in the sample (not just neighboring points as in the two output case). In this case, D is defined as the euclidean distance between points on the surface of the triangle.

Then the Parzen density function for a point $\mathbf{x}$ is defined as :

$$F(\mathbf{x}) = \sum_{i=1}^{N} \frac{\exp(-((\mathbf{x}-\mathbf{o}_i)^2 / 2\sigma^2))}{\sigma\sqrt{2\pi}}$$

Where N is the number of neural networks generated by the Monte Carlo simulation, $\mathbf{o}_i$ is the output of the i[th] network, and σ is the smoothing constant.

Then I define a grid inside the triangular surface. This is formed by three sets of N equally spaced lines at regular intervals parallel to each side of the triangle. This tiles the space with $(N + 1)^2$ subtriangles where N is the number of grid lines per side.

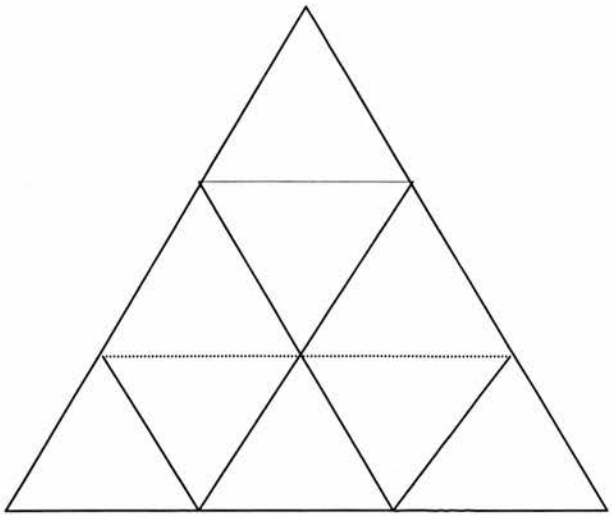


Figure 1: Triangular grid with $N = 2$

Throughout this thesis I have used $N = 79$. Then I use $F(\boldsymbol{x})$ to calculate the height at each grid vertex. The height of the surface is scaled so that the volume equals one. In order to display the surface graphically (e.g. see figure 20, chapter 2), I have drawn the level contours for a series of confidence regions. The confidence levels I have chosen are 90%, 70%, 50%, and 25%. To draw these, I sort the subtriangles in the grid by volume. Then starting with the biggest volume, I count out subtriangles accumulating their volume until I reach the desired percentage. Then I draw a boundary (or set of boundaries if the distribution has multiple modes) around the selected subtriangles. Thus, within the limits of the discrete approximation, the contour for the $N\%$ confidence region is the level contour that contains $N\%$ of the probability.

A.2 Beta and Dirichlet Functions

In order to validate the density estimates described above, I have compared them in simple cases to Beta and Dirichlet functions (sections 2.6 and 2.12), which can be

computed to arbitrary precision. To do this I have used the same approach based on discrete grids as described above for density estimation, but I have replaced the Parzen function with an exact calculation of probability. The idea is that we run a series of trials with two or three possible outcomes (outputs). The observed (input) values are fixed for the duration of the trial. Therefore we can replace the Parzen window function with a function that calculates the relative likelihood of various estimates of the true probability distribution over outcomes based on the finite series.

A.2.1 Two Output Case (Beta Function)

If we have N trials with M positive results, then we can estimate the likelihood based on the series that the true probability of a positive result is P as

$$\mathrm{F}(P) = \binom{N}{M} P^{M} (1-P)^{N-M}$$

Again, the likelihood is computed on a discrete grid, and the curve is scaled so that the area under the curve equals one.

A.2.2 Three Output Case (Dirichlet Function)

Say the outcome of a trial can have three values : $\mathbf{o}_1, \mathbf{o}_2$ and $\mathbf{o}_3$
If we have N trials resulting in $\mathbf{M}_i$ observations for each of the three $\mathbf{o}_i$, then we can compute the likelihood based on the series that the true probability distribution we should assign to our output classes $\mathbf{o}$ is $\mathbf{P}$ as

$$\mathrm{F}(\mathbf{P}) = \binom{N}{\mathbf{M}_1}\binom{N-\mathbf{M}_1}{\mathbf{M}_2} \mathbf{P}_1^{\mathbf{M}_1} \mathbf{P}_2^{\mathbf{M}_2} \mathbf{P}_3^{\mathbf{M}_3}$$

Again, the likelihood is computed on a discrete triangular grid (see A.1.2 above), and the surface is scaled so that the volume under the surface equals one.

A.3 Parameters for the simulations

A.3.1 Neural Network Model Specifications

Hyperprior Widths are the width parameters for the gamma distribution, and the Hyperprior Alphas are the shape parameters

Two Class Problems

Input Nodes:	2
Hidden Nodes:	8
Output Nodes	2
Activation Function	Softmax
Automatic Relevance Detection	No
Input To Hidden Hyperprior Width	2.0
Input To Hidden Hyperprior Alpha	5.0
Hidden Bias Hyperprior Width	0.5
Hidden Bias Hyperprior Alpha	5.0
Hidden To Output Hyperprior Width	0.25
Hidden To Output Hyperprior Alpha	2.5
Output Bias Hyperprior Width	0.25
Output Bias Hyperprior Alpha	2.5

Three Class Problems

Input Nodes	varied
Hidden Nodes:	12
Output Nodes	3
Activation Function	Softmax
Automatic Relevance Detection	No
Input To Hidden Hyperprior Width	4.0
Input To Hidden Hyperprior Alpha	10.0
Hidden Bias Hyperprior Width	1.0
Hidden Bias Hyperprior Alpha	10.0
Hidden To Output Hyperprior Width	0.5
Hidden To Output Hyperprior Alpha	5.0
Output Bias Hyperprior Width	0.5
Output Bias Hyperprior Alpha	5.0

A.3.2 Monte Carlo Markov chain Specifications

These are the same for all simulations:

Initial Phase

The commands to Radford Neal's software would be:
mc-spec <file name> repeat 10 heatbath hybrid 100:10 0.2
net-mc < file name > 1

My wrapper runs off a command file that would contain these commands:

```
mcmc-iterate <file name> 1
 repeat 10
  heatbath
  hybrid 100:10 0.2
 end
end
```

This means that we have an inner loop that runs 10 times alternating Gibbs Sampling (heatbath) updates that replace the momentum terms with hybrid Monte Carlo with 100 leapfrog steps using a window of 10 and a stepsize adjustment factor 0f 0.2. The inner loop is contained in an outer loop that runs once. The specified file contains the specification for the neural network.

Sampling Phase

The commands to Radford Neal's software would be:
mc-spec <file name> repeat 10 heatbath 0.95 hybrid 100:10 0.3
net-mc < file name > 120

My wrapper runs off a command file that would contain these commands:

```
mcmc-iterate <file name> 120
 repeat 10
  heatbath 0.95
  hybrid 100:10 0.3
  negate
 end
end
```

The semantics are the same as for the initial phase.

References

Ambroso, C., Bowes C., Carson E., et al., 1992, INFORM: development of information management and decision support systems for High Dependency Environments, *Journal of Clinical Monitoring and Computing*, 8: 295-301

Andersen, H..C. 1980. Molecular dynamics simulations at constant pressure and/or temperature, *Journal of Chemical Physics, vol. 72, 2384-2393*

Andrews, P.J.D., McQuatt A., Jones P.A., et al.., 1999. Decision tree analysis of demographic, time series physiological and medical complication data after traumatic brain injury, *British Journal of Anaesthesia* 82(2): 455-456

Andrews P.J.D., Muruguval S., Deeham S., 1996. Conventional multimodality monitoring and failure to detect ischemic cerebral blood flow, *Journal of Neurosurgical Anaesthesiology* 8: 220-226

Barlow P.G., Teasdale G., Jennett B., et al. 1984. Computer assisted prediction of outcome of severely head-injured patients, *Journal of Microcomputer Applications*, 7; 271-277

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

Bishop, C.M., 1995, *Neural Networks for Pattern Recognition,* Clarendon Press, Oxford

Bishop, C.M. (ed.), 1998. *Neural Networks and Machine Learning.* Springer

Black, N., 1996. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, 312; 1215 - 1219, May 1996

Breiman,L., 1996. Bagging Predictors, *Machine Learning*, 26(2). 13-140

Braakman, R., Gelpke G.J., Habbema J.D.F. , et al., 1980. Systematic selection of prognostic features in patients with severe head injury. *Neurosurgery*. 6: 362-370

Chan, K.H., Miller J.D., Dearden N.M., et al., 1992, The effect of changes in cerebral perfusion pressure upon middle cerebral artery blood flow velocity and jugular bulb venous oxygen saturation after severe brain injury., *Journal of Neurosurgery* 77: 55-61

Coiera, E., 1991. Medical informatics, *British Medical Journal*, 310: 1381-1387

Cormio, M., Valadka A.B., a nd Robertson C.S., Elevated jugular venous oxygen saturation after severe head injury. 1999. *Journal of Neurosurgery* (90) 9-15

Corrie J., Piper I.R., Housley, A., et al., Microcomputer based data recording: Improved indentification of secondary insults in head injury patients., 1993. *British Journal of Intensive Care.* 6: 225-233

Cortes C., and Vapnik V. 1995. Support Vector Network, *Machine Learning,* (20) 273-297

Cristianini N., and Shawe-Taylor J., 1999. Bayesian voting schemes and large margin classifiers. In Schoelkopf B., Burges C. and Smola A. (eds.) *Advances in kernel Methods - Support Vector Learning, 55-68*, MIT Press

Czonyka, M. et al. 1994. *Journal of Clinical Monitoring and Computing*, 11: 223-232.

Duane, S., Kennedy, A.D., Pendleton B.J., Roweth, D. 1987. Hybrid Monte Carlo, *Physics Letters B,* 195, 216-222

Duda, R.O. and Hart P.E., 1973. *Pattern Classification and Scene Analysis*, John Wiley

Dybowski, , R., and Weller P., 1996. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm., *Lancet:* vol 347, issue 9009, pg. 1146.

Hart A. and Wyatt J., 1990. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. *Medical Informatics.* 15: 229-236

Fackler, J.C., 1998. From data to information: multiparameter alerts. *International Journal of Intensive Care.* Summer 1998 supplement

Feng C., and Michie D., Machine learning of rules and trees, 1994. in *Machine Learning, Neural and Statistical Classification*, Michie D., Spiegelhalter D.J., and Taylor C.C. (eds.), Ellis Horwood Ltd.

Foulkes M.A., Eisenberg H.M., Jane J.A., et al., 1991, The Traumatic Coma Data Bank: design, methods, and baseline characteristics, *Journal of Neurosurgery,* 75; S8-S13

Gopinath S.P., Robertson, C.S., Contant C.F., et al., 1994. Jugular venous desaturation and outcome after head injury. *Journal of Neurosurgical Psychiatry,* 57; 717-723

Holland, J.H. 1986. Escaping brittleness: the possibilities of general purpose learning algorithms applied to parallel rule-based systems. In Michalski R.S., Carbonell J.G. and Mitchell T.M. (eds.), *Machine Learning*, vol. 2, 593 - 623. Morgan Kaufmann

Howells, T.P., Piper I.R., Jones P.A., Souter, M., and Miller, J.D., Design of a research database for the study of secondary insults following head injury. 1995. *Journal of Neurotrauma* 12(3), 471

Jennett, B., Bond, M.. 1975. Assessment of outcome after severe brain damage. *Lancet* i; 480

Jennett, B., Teasdale G., Galbraith S., et al., 1977. Severe head injuries in three countries, *Journal of Neurology, Neurosurgery and Psychiatry*, 40;291-298

Jones, P.A., Andrews, P.J.D., Midgley S., et al., 1994, Measuring the burden of secondary insults in head-injured patients during intensive care. *Journal of Neurosurgical Anesthesiology* 6(1): 4-14

Lang, K.J. and Witbrock, M.J..1988, Learning to tell two spirals apart, *Proceedings of the 1988 Connectionist Models Summer School*, Morgan Kaufmann.

MacKay, D.J.C., 1992a, Bayesian interpolation, *Neural Computation*, 4(3), 415-447

MacKay, D.J.C., 1992b, A practical Bayesian framework for backpropagation networks, *Neural Computation*, 4(3), 448-472

MacKay, D.J.C., 1992c, The evidence framework applied to classification networks, *Neural Computation*, 4(5), 720-736

MacKay, D.J.C., 1993, Bayesian Non-linear modeling for the Energy Prediction Competition, *ASHRAE Transactions*, vol. 100, part 2

MacKay, D.J.C. 1994. Hyperparameters: optimise or integrate out? In *Maximum Entropy and Bayesian Methods*. Heidbreder, G. (ed.), Kluwer

MacKay, D..J.C., 1998, Introduction To Gaussian Processes, in (Bishop, C.M. (ed.), 1998)

Macmillan C.S.A., Andrews P.J.D., Jones P.A., et al., 1998, Poor outcome associated with elevated jugular bulb saturation in acute brain injury. *Neuro-anaesthesia Societies of Finland, Scandinavia, Great Britain, and Ireland,* Helsinki, 1998.

Marmarou, A., Anderson S.L., Ward J.D., et al., 1991, Impact of ICP instability and hypotension on outcome of patients with severe head injury. *Journal of Neurosurgery*; 75 S59 - S66

Marion, D.W., Carlier P.M., 1994. Problems with initial Glasgow Coma Scale assessment caused by prehospital treatment of patients with head injuries: results of a national survey. *Journal of Trauma,* 36(1)

Michie, D., Spiegelhalter D.J., and C.C. Taylor (eds.), 1994. *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Ltd.

Miller. J.D., 1992. Evaluation and treatment of head injury in adults. *Neurosurgery Quarterly.* 2(1): 28-43

Neal, R.M., 1996, *Bayesian Learning for Neural Networks*, Springer-Verlag, New York

Neal R.M., 1993. Probabalistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto

Piper. I.R., Lawson A., Dearden N.M., and Miller J.D., 1991. Computerised data collection: a microcomputer data collection system in head injury intensive care. *British Journal of Intensive Care.* May/June 1991

Piper, I.R., Contant C.F., Citerio G., et al. 1999. Multi-centre assessment of the Spiegelberg compliance monitor. *International Meeting on Brain Oedema,* Newcastle

Persson, L., Hillered L., 1992. Chemical monitoring of surgical intensive care patients using intracerebral microdialysis. *Journal of Neurosurgery.* 76; 72-80

Ripley, B.D., 1996. *Pattern Recognition and Neural Networks,* Cambridge University Press

Ripley, B.D., 1998 Statistical theories of model fitting. In Bishop (ed.) 1998

Rosner M.J., 1985. The vasodilatory cascade and intracranial pressure. *Intracranial Pressure VI*, Springer Verlag., 137 - 141

Rumelhart D.E., Hinton G.E., and Williams R.J., 1986, "Learning internal representations by back-propagating errors", in (Rumelhart, McClelland, et al., 1986)

Rumelhart D.E., McClelland J.L., and the PDP Research Group. 1986. *Parallel Distributed Processing Explorations in the Microstructure of Cognition,* vol. 1, MIT Press.

Schapire, R.E., 1999. A brief introduction to boosting. *Proceedings of the 16th International Joint Conference on Artificail Intelligence*

Sedbrook T., Wright W., and Wright R., 1991. Application of a genetic classifier for patient triage, *Fourth International Conference on Genetic Algorithms,* Morgan Kaufmann

Signorini, D.F., Andrews, P.J.D., Jones P.A., et al., 1999a, Predicting survival using simple clinical variables: a case study in traumatic brain injury. *Journal of Neural Neurosurgical Psychiatry*, (66) 22-25

Signorini, D.F., Andrews, P.J.D., Jones, P.A. et al., 1999b. Adding insult to injury: the prognostic value of early secondary insults for survival after traumatic brain injury. *Journal of Neural Neurosurgical Psychiatry*, (66) 26-31

Teasdale G., Jennett B., 1974. Assessment of coma and impaired consciousness. *Lancet* 2:81

Teasdale G., 1981. Prognosis after severe head injury, in *Management of Head Injuries*, B. Jennett and G. Teasdale (eds.), S.A. Davies & Co., Philadelphia, Ch. 14: 317-332

Titterington, D.M., Murray G.D., Murray, L.S., et al., Comparison of discrimination techniques applied to complex data set of head injury patients. 1981. *Journal of the Royal Statistical Society*, 144 145-175

Vapnik, V. 1998. The support vector method of function estimation, in (Bishop, C.M. (ed.), 1998)

Vespa, P.M., Nuwer M.R., Csaba J., et al., 1997., Early detection of vasospasm after acute subarachnoid hemorrhage using continuous EEG ICU monitoring. *Electroencephalography and clinical Neurophysiology*; 103; 607-615

Weigend and Gershenfield (eds.), 1994, *Time Series Prediction: Forecasting the future and understanding the past*, Proceedings Volume XV in the Santa Fe Institute Studies in the Sciences of Complexity series, Addison Wesley

Wolpert, D.H., 1993, On the use of evidence in neural networks, in *Advances in Neural Information Processing Systems,* Hanson S.J., Cowan, J.D., and Giles C.L. (eds.), Morgan Kaufmann

Wolpert, D.H., 1994, Bayesian backpropagation over I-O functions rather than weights, In *Neural and Information Processing Systems*, Morgan Kaufmann, San Francisco