SEVERITY OF ILLNESS SCORING IN THE INTENSIVE CARE UNIT: A COMPARISON OF LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS

By

Gordon S. Doig
Graduate Program in Epidemiology
and Biostatistics

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario
April, 1999

THE UNIVERSITY OF WESTERN ONTARIO
FACULTY OF GRADUATE STUDIES

CERTIFICATE OF EXAMINATION

Chief Advisor	Examining Board
Dr. J. McD. Robertson	
Advisory Committee	
Dr. C. M. Martin	
Dr. W. J. Sibbald	
Dr. D. B. Chalfin	

The thesis by

Gordon S. Doig

entitled

Severity of illness scoring in the intensive care unit: A comparison of logistic regression and artificial neural networks.

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Date _____ _____

Chair of Examining Board

Abstract

Purpose: To compare the predictive performance of a series of logistic regression models (LMs) to a corresponding series of back-propagation artificial neural networks (ANNs).

Location: A 30 bed adult general intensive care unit (ICU) that serves a 600-bed tertiary care teaching hospital.

Patients: Consecutive patients with a duration of ICU stay greater than 72 hours.

Outcome: ICU-based mortality.

Methods: Data were collected on day one and day three of stay using a modified APACHE III methodology. A randomly generated 811 patient developmental database was used to build models using day one data (LM1 and ANN1), day three data (LM2 and ANN2) and a combination of day one and day three data (LM_{OT} and ANN_{OT}). Primary comparisons were based on area under the receiver operating curves (aROC) as measured on a 338 patient validation database. Outcome predictions were also obtained from experienced ICU clinicians on a subset of patients.

Results: Of the 3,728 patients admitted to the ICU during the period from March 1, 1994 through February 28, 1996, 1,181 qualified for entry into the study. There was no significant difference between LM and ANN models developed using day one data. The ANN developed using day three data performed significantly better than the corresponding LM (aROC LM2 0.7158 vs. ANN2 0.7845, p=0.0355). The time dependent ANN model also performed significantly better than the corresponding LM (aROC LM_{OT} 0.7342 vs. ANN_{OT} 0.8095, p=0.0140).

The predictions obtained from ICU consultants (aROC 0.8210) discriminated significantly better than LM_{OT} (aROC 0.6814, p=0.0015) but there was no difference between the consultants and ANN_{OT} (aROC 0.8094, p=0.7684).

Conclusion: Although the 1,181 patients who became eligible for entry into this study represented only 32 percent of all ICU admissions, they accounted for 80 percent of the resources (costs) expended. ANNs demonstrated significantly better predictive performance in this clinically important group of patients. Four potential reasons are discussed: 1) ANNs are insensitive to problems associated with multicollinearity; 2) ANNs place importance on novel predictors; 3) ANNs automatically model nonlinear relationships and; 4) ANNs implicitly detect all possible interaction terms.

Keywords: intensive care, critical care, severity-of-illness, logistic regression, artificial neural networks, genetic algorithms, back-propagation, receiver operating characteristic, predictive model building

Table of Contents

List of Tables

List of Figures

List of Equations

List of Appendices

List of Abbreviations

1. Introduction

In North America, the intensive care unit (ICU) accounts for seven percent of all hospital beds, 15 to 20 percent of all hospital expenditures, and approximately one percent of the Gross National Product.[1] Because the demand for intensive care is growing and resources are increasingly constrained,[2] it is becoming even more important to make effective decisions with respect to management practices, utilization, and even individual therapy received in the ICU. It has long been recognized in the field of intensive care medicine that analytical epidemiology can play a major role in providing tools for ICU-level management and decision support.[3,4,5,6]

Predictive modeling can provide an estimate of the risk of mortality faced by a patient upon entry into the ICU. This risk of mortality, commonly expressed as a severity-of-illness score, can serve to support quality assurance activities, resource allocation, and the evaluation of novel therapies.[7,8,9] Severity-of-illness scores usually perform well when used to predict the expected overall mortality experience of an entire ICU population. Research has shown however, that as a patient's length of stay in the ICU increases, the accuracy of admission scores decreases drastically. Severity-of-illness scores also consistently fall short of clinical usefulness when used to predict mortality in individual patients.[8,10,11]

Some researchers have suggested that gaining a better understanding of the complex patterns associated with mortality prediction in the ICU will require the use of novel mathematical approaches, such as set theory or fuzzy logic.[12,13] Artificial neural networks (ANNs) represent an alternative approach to modeling complexity in the ICU. ANNs are a group of techniques developed by cognitive scientists to model the human brain's methods of learning, have also been shown to be useful in situations such as the ICU where the relationships between variables and outcomes are complex.[14]

In the engineering disciplines of image processing, speech recognition and natural language processing, neural network techniques have demonstrated the ability to outperform classical statistical methods.[15] In the domain of artificial intelligence (AI), research has shown that neural networks can easily solve many complex problems which conventional AI systems find difficult or impossible to solve.[16] In medicine, neural

networks have outperformed clinicians on the diagnosis of hepatic masses, pulmonary emboli and breast tumors.[17, 18, 19] Although studies in the medical literature have compared neural networks against other image processing techniques and against clinicians, to date few published studies have used objective statistical measures to compare the ability of ANNs to predict patient outcomes against currently accepted analytical techniques.

The purpose of this project was to compare the performance of artificial neural networks with that of multivariate logistic regression in predicting mortality in patients admitted to the intensive care unit. The potential clinical utility of this novel method was also evaluated by comparing the predictions of the artificial neural network with those of senior clinicians. In today's arena of advancing medical technology and retreating budgets, an improved decision support tool would be invaluable to clinicians, unit managers and ultimately, to the patients themselves.

2. Literature Review

2.1 Mortality Prediction in the Intensive Care Unit

Many different scoring systems are used to predict the risk of mortality experienced by adult ICU patients. The most widely used and well validated include the Acute Physiology and Chronic Health Evaluation (APACHE), the Mortality Probability Models (MPMs) and the Simplified Acute Physiology Score (SAPS).[7,20,21] While the specific methodology for data collection and risk estimation for each of these systems differ markedly, there are some striking basic similarities. They all combine certain measures of physiological status and/or treatment modalities with pre-existing risk factors to produce a logistic regression-based measure of risk of mortality. Independent validation studies comparing risk predictions from these three systems have consistently failed to demonstrate any notable differences in performance.

2.2 The APACHE Scoring System

The original APACHE scoring system, which utilized information readily available from the medical record, was first introduced in 1981.[22] In subsequent studies, it was reported to be a reliable and valid classification system for critically ill patients.[23,24] An excellent detailed overview of its early development is available elsewhere.[25]

The APACHE scoring system was first developed on a database of 804 consecutive ICU admissions. The study protocol called for recording 34 separate physiological variables during the first 36 hours of ICU admission. A score between zero and four was assigned to each physiologic value based on its deviation from normal and these individual scores were then summed to produce the overall patient score. The selection of the 34 physiologic variables and the score attributed to each was determined by expert consensus.[22]

2.2.1 APACHE II

Driven by the need to develop *and* validate a scoring system on a representative database, the development of APACHE II was undertaken.[26] APACHE II was developed and validated on a database of 5,815 ICU admissions from 13 hospitals and reduced the

number of required admission acute physiology score (APS) variables from 34 to12 by dropping infrequently measured variables.

The twelve physiologic variables that comprise the APACHE II model are: temperature, mean arterial pressure, heart rate, respiratory rate, oxygenation, arterial pH, serum sodium, serum potassium, serum creatinine, hematocrit, white blood count, and the Glasgow coma scale score. For each variable, 'points' were assigned based on the worst value (the value deviating farthest from normal) recorded over the first 24 hours of ICU admission. Individual variable points were summed to form the APS component of the APACHE II score.

The complete APACHE II score also included points awarded for age and for the chronic health evaluation (CHE) variables. The CHE awarded points for chronic liver, cardiovascular, respiratory, renal and immunological failure. The APS, age and CHE points were then summed and offered to a logistic regression model to calculate a predicted risk of mortality. The points awarded in the APS, age and CHE components of the APACHE II model were based on an expert consensus process.

The reliability of information abstracted from medical records for APACHE II has been extensively measured.[27] Abstraction of the APS component was found to have an intraclass correlation coefficient (ICC) of 0.90. The abstraction of age information was found to have an ICC of 0.99 and the ICC for the reproducibility of the CHE component was 0.66. The investigators concluded that the collection of information for the calculation of the APACHE II score was highly reliable.

APACHE II has been used to compare utilization rates and outcomes of critical care services throughout North America, Hong Kong, Switzerland, New Zealand, Japan and most recently, Tunisia.[28,29,30,31,32,33] In 1995 an extensive evaluation was undertaken comparing 1,724 consecutive admissions at two Canadian ICU's with data collected on 4,087 consecutive admissions in 13 ICU's based in the United States.[34] In this study, both the overall mortality experience (24.8% vs. 22.1%, p=0.028) and the average admission APACHE II score (16.5±0.2 vs. 14.8±0.1, p=0.0001) were significantly higher in the Canadian ICUs. Furthermore, by using a graphical technique, the authors demonstrated that when observed mortality rates were controlled for severity of illness using APACHE II scores, no observable difference existed between the Canadian and American ICU's. In

this study, the area under the receiver operating characteristic (aROC) curve for the APACHE II model in the Canadian ICUs was 0.86. The ROC curve is simply a graph of the sensitivity vs. (1-specificity) for each possible cutoff value. The area under the ROC curve serves as a useful measure of discrimination. It represents the proportion of all possible pairs of outcomes in which the predicted risk of mortality for the patient who actually died is higher than the predicted risk of mortality for the patient who lived.[35]

Although the APACHE II system has been used extensively for 'benchmarking' performance between countries, some reports have suggested that both calibration and discrimination of APACHE II decreases when applied to patient populations that were not included in the original development project.[33,36]

2.2.2 APACHE III

The development of the APACHE III prognostic scoring system is outlined in detail in a comprehensive series of articles.[1,37,38,39,40,41,42] It was developed from an ICU database of 17,457 consecutive patients collected from 40 hospitals throughout the United States. Of these 40 hospitals, 26 volunteered to participate as a result of a random selection process and 14 non-randomly selected hospitals volunteered. The APACHE III database is considered to be representative of American ICUs.

In addition to the original 12 variables of the APACHE II, the APS component of the APACHE III contained five additional physiological variables: blood urea nitrogen (BUN), urine output, serum albumin, bilirubin, and glucose. Overall explanatory power of the APACHE III also improved when the following interactions were considered: serum pH with $PaCO_2$, serum creatinine with urine output, and respiratory rate with ventilator use.

The comorbid disease states considered as important predictors of outcome by APACHE III included: acquired immunodeficiency syndrome, hepatic failure, lymphoma, solid tumor with metastasis, leukemia/multiple myeloma, immunocompromise, and cirrhosis.[9] Of all the comorbid chronic health states, the only ones to meet the statistical requirements for inclusion in APACHE III were those that influence the patient's immunologic status.

The overall predictive power of APACHE III logistic regression risk estimates developed on the data obtained during the first day of ICU stay is evidenced by the total

model r^2 of 0.41 and area under the ROC curve of 0.90. Overall correct classification on the first day was 88.2 percent. These values were reported as being an improvement over both the area under the ROC curve (0.85) and the overall correct classification rate (85.5%) of APACHE II.[9]

In a recent study looking at 200 non-selected head trauma patients conducted in China, the discrimination of APACHE III, as measured by area under the ROC curve, was reported to be 0.90; whereas that of APACHE II was 0.84.[43] In this relatively small study, no statistically significant difference was evident between the performance of APACHE II and APACHE III. However, a large independent evaluation of APACHE III conducted on 14,745 consecutive admissions to 137 ICUs throughout Europe and North America, reported that APACHE III had significantly improved predictive performance over APACHE II (aROC APACHE III 0.866 vs. aROC APACHE II 0.853, $p<0.0001$).[44]

In 1996, an evaluation of APACHE III was undertaken in a Brazilian patient population.[45] In this article, the APACHE III predicted mortality rate of 20 percent was significantly lower than the actual 34 percent mortality rate reported in the 1,734 patient cohort ($p<0.0001$). The area under the ROC curve was reported as 0.82.

Further validation of the APACHE III score was undertaken in a consecutive sample of 37,668 ICU admissions accumulated from 1993 to 1996 from 285 non-randomly selected ICUs located in 161 American hospitals.[46] This study reported that although no significant difference existed between the aggregate observed mortality rate and that predicted by the APACHE III model (12.35% vs. 12.27%, $p=0.541$), the Hosmer-Lemeshow goodness of fit (H-L gof) test displayed a significant lack of calibration (χ^2=48.71,8df(*sic*);$p<0.0001$). The aROC was reported as being 0.89, which was considered good for a validation study.

Most recently, in order to improve predictive performance in a cohort of Spanish patients, the logistic regression coefficients for the APACHE III equations were recalculated.[47] By contacting the directors of all ICUs in the Spanish National Health Service hospital network, data were collected on 10,929 patients from 86 participating ICUs. These data were abstracted from patient charts using the APACHE III

methodology and the logistic regression coefficients for the APACHE III risk prediction model were recalculated based on this Spain-specific database.

The area under the ROC curve for the recalibrated APACHE III equations was reported as 0.83 on the developmental database and 0.82 in the 'cross-validated model'. Although the Hosmer-Lemeshow goodness of fit statistic was reported as being non-significant [χ^2=12.27,10df(*sic*);p=NS] on the developmental database, no goodness of fit tests were performed on the cross-validated model.[48]

2.3 The Mortality Probability Model (MPM) Scoring System

In 1985, a multiple logistic regression (MLR) model was developed to predict outcome in ICU patients based on 755 consecutive admissions to the medical/surgical ICUs at Baystate Medical Center, Springfield, Massachusetts.[49] The major objective of developing this MLR was to create a severity-of-illness score using an objective methodology to assign variable weights.

A total of 137 background, condition, and treatment variables were collected at admission to the ICU. Seventy-five of these variables were recollected at 24 and 48 hours from patients still in the ICU. Tests of association of each study variable with vital status at hospital discharge were performed using the Student's *t*-test with continuous variables, and the Chi-square test of independence with categorical variables.

Of the 137 admission variables assessed, the only ones that were significantly related to outcome using univariate analysis were: age, systolic blood pressure, heart rate, number of organ failures, source of admission, presence of infection, cardiopulmonary resuscitation before admission, type of admission (elective/emergency), PaO_2, bicarbonate, serum creatinine and level of consciousness (coma or deep stupor vs. other). A forward stepwise process was then used to develop a multivariate model.

The multivariate model demonstrated good fit (H-L gof p=0.3871). Graphical representations of the ROC curves were presented but the area under the curves was not calculated. The performance of this initial model was not evaluated on a validation database.

Subsequently, the results of a more extensive undertaking were published.[50] In this second paper by the same investigators, data were presented on 2,783 admissions to

the Baystate Medical Center and four distinct mortality probability models (MPMs) were created. The first model was based on variables collected at the time of admission to the ICU (MPM_0) and two subsequent models were based on data available at 24 hours (MPM_{24}) and 48 hours (MPM_{48}) of stay in the ICU. The fourth model, termed MPM over time (MPM_{OT}), was developed using only three independent variables: X_1=pred(MPM_O), X_2=pred(MPM_0)-pred(MPM_{24}), and X_3=pred(MPM_{24})-pred(MPM_{48}) where pred(MPM) represents the predicted probability of mortality from a particular MPM model. All models displayed good calibration based on the Hosmer-Lemeshow goodness of fit. Formal measures of discrimination were not reported and the performance of the models was not evaluated on an independent validation database.

2.3.1 MPM II

The development of the second-generation mortality probability model (MPM II) was based on a database of 19,124 patients from 137 medical/surgical ICUs in 12 different countries.[51] Of the 19,124 available patients, 12,610 were used to develop the model and 6,514 were retained in a separate independent database for model validation.

In this paper, two unique MPM II models were created: MPM II_0 was based on information available at entry to the ICU and MPM II_{24} was based on information abstracted after 24 hours of stay. For both models, inclusion of main effects in the multivariate model was based on a univariate probability less than 0.10. All possible two-way interactions were investigated and retained in the model if: 1) they demonstrated statistical significance ($p<0.05$) in the full model; 2) the combination of factors from the main effects contributing to the model (a AND b) was present in at least one percent of the population; and 3) the interaction satisfied the requirement of 'clinical plausibility'. No two-way interactions satisfied all three criteria for either MPM II_0 or MPM II_{24}.

MPM II_0 contained 15 main effect terms and exhibited good performance on both the developmental and validation database (H-L gof p=0.62, aROC=0.84 developmental, H-L gof p=0.33, aROC=0.82 validation). The MPM II_{24} contained 13 variables and also performed well (H-L gof p=0.76, aROC=0.84 developmental, H-L gof p=0.23, aROC=0.84 validation).

In a separate study, two additional MPM II models were developed: MPM II_{48} required information available 48 hours and MPM II_{72} required information available 72

hours post-admission. This study was based on a database of 6,290 patients admitted to six adult medical and surgical ICUs in Massachusetts and New York State. Of the total 6,290 available patients, 3,023 had complete data available at 48 hours and 2,233 had complete data available at 72 hours post-admission.[52]

Both the MPM II_{48} and the MPM II_{72} contained the same 13 variables and coefficients as the MPM II_{24}. Only the intercept term was recalculated to reflect "the increasing probability of mortality with increasing length of stay in the ICU".[52] Both the MPM II_{48} (H-L gof p=0.31, aROC=0.81 developmental, H-L gof p=0.59, aROC=0.80 validation) and the MPM II_{72} (H-L gof p=0.31, aROC=0.79 developmental, H-L gof p=0.41, aROC=0.75 validation) displayed good performance.

In 1994, the performance of MPM II_0 was assessed on an independent database composed of 8,724 admissions to 26 ICUs in Britain and Ireland.[36] When applied to this database, MPM II_0 displayed poor fit (H-L gof $p<0.0001$) with a reported area under the ROC curve of 0.74. Evaluation in a smaller 1,325 patient database from three Tunisian ICUs revealed that both MPM II_0 and MPM II_{24} displayed poor fit (H-L gof $p<0.001$ for both).[33] The area under the ROC curve was reported as 0.85 for MPM II_0 and 0.88 for MPM II_{24} in this database. In a 1998 publication, MPM II_0 also displayed poor fit (H-L gof $p<0.0001$) when applied to an independent database of 10,027 evaluable patients collected from consecutive admissions to 89 ICUs in 13 European countries.[53] The area under the ROC curve in this validation study was 0.78 with a standard error of 0.006.

2.4 Simplified Acute Physiology Score (SAPS)

The original SAPS scoring system was developed as an extension of the first APACHE model. [54] The objective of SAPS was to simplify the APS portion of APACHE by using expert consensus to reduce the number of variables that would be abstracted from each patient chart. The SAPS variables were collected using the same methodology developed by APACHE (worst value over first 24 hours) and SAPS used a similar subjective point allocation scheme as APACHE. To produce the SAPS score, points were allocated for each variable's deviation from normal and summed. The original SAPS did not attempt to produce a prediction of mortality for a given score and thus was simply a 'scoring' system composed of 13 physiologic variables plus age.

2.4.1 SAPS II

The second generation SAPS model (SAPS II) was developed on a 13,152 patient database collected from 137 medical/surgical ICUs in 12 countries.[55] This database was randomly divided into developmental (65 percent) and validation (35 percent) data sets.

SAPS II considered all the original variables comprising SAPS for eligibility plus additional demographic and physiologic variables. Inclusion in the final score was considered if a variable demonstrated a significant univariate relationship with mortality or contributed towards improved fit in the final model. The final SAPS II model contained 17 variables: 12 physiological variables (temperature, heart rate, blood pressure, white blood cell count, bilirubin, serum sodium, serum potassium, serum bicarbonate, blood urea nitrogen, urine output, oxygenation and a measure of neurological status) and 5 demographic variables (age, type of admission, presence of AIDS, presence of a hematologic malignancy, and presence of metastatic cancer).

Points awarded to each variable range were determined using an objective approach and summed to calculate the SAPS II score. To translate the SAPS II score into a predicted probability of mortality, the entire score would be entered into the SAPS II logistic equation. Although the SAPS II predictive equation contains only the SAPS II score, it demonstrated good performance in both the developmental (H-L gof $p=0.88$, aROC=0.88) and validation (H-L gof $p=0.10$, aROC=0.86) data sets.

The performance of SAPS II has been evaluated in independent data sets. In a consecutive sample of 1,325 patients from three Tunisian ICUs, SAPS II was shown to have poor fit (H-L gof $p<0.001$) but good discrimination (aROC=0.84).[33] In a larger study involving 10,027 patients collected from 89 ICUs in 13 European countries, SAPS II demonstrated poor fit (H-L gof $p<0.001$) with an area under the ROC curve of 0.82 and a standard error of 0.005. In this study, the area under the ROC curve for SAPS II was found to be significantly better than the area reported for the MPM II_0 model, which was evaluated in the same study (0.82 ± 0.005 vs. 0.79 ± 0.006, $p<0.001$).[53]

The performance of SAPS II has also been evaluated in an intermediate care unit. The requirement for intermediate care is defined by patients who are not currently suffering from a life threatening condition, who are not currently receiving invasive interventions but who do require intensive monitoring. Although SAPS II was not

developed in intermediate care units, a large portion of the original SAPS II development database contained patients who satisfied the requirements for intermediate care. In this study of 561 consecutive admissions to a French intermediate care unit, 433 patients qualified for SAPS II scoring. Based on these 433 patients, SAPS II displayed good calibration (H-L gof $p>0.5$) and discrimination ($aROC=0.85\pm0.04$).[56]

2.5 Day of Severity-of-Illness Scoring

2.5.1 APACHE

The original intent of the APACHE scoring system was to use data available shortly after admission to the ICU to aid in assessment of a patient's risk of mortality. In subsequent research, it was recognized that between the first 24 to 48 hours after admission to the ICU, severely ill patients often develop serious sequelae such as line infections, ventilator associated pneumonia, sepsis, refractory shock and/or acute respiratory distress syndrome. The development of these complications significantly alters a patient's risk of mortality and has been shown to be more strongly associated with outcomes than are the APACHE II admission scores.[11]

In an attempt to understand the performance of the APACHE II scoring system in patients with lengths of stay greater than 72 hours, a study of 110 consecutive patients demonstrated that the predictive accuracy of the APACHE II system decreased with the length of time the patient stayed in the ICU.[57] Furthermore, it was also shown that if APACHE II scores were recalculated on each day of stay and scores remained high in the face of continued maximal intervention, fatal outcomes could be accurately predicted.[13]

In 1994, the methodology of the APACHE III risk prediction system was formally extended to incorporate repeated physiologic measures over time in order to improve its predictive accuracy.[58] The APACHE III database, which included 17,440 patients collected from consecutive admissions to 42 ICUs at 40 different hospitals, was used to develop a unique predictive equation for each day the patient remained in the ICU.

The basic form of the terms entered into these logistic regression equations was: Daily Risk = (APS day 1) + (APS current day) + (change in APS since yesterday). The equations also contained the following non time-dependent variables: indication for ICU admission, location and length of treatment before ICU admission, patient's age and CHE

score. These variables were not selected for entry into the model based on objective statistical criteria, but were pre-specified by a panel of experts. The odds ratios and areas under the ROC curves for each daily model are reported in Table 1.

Table 1. Odds ratios and area under the ROC curve for the APACHE III daily risk estimate models.[58]

		Odds Ratios			
ICU Day of stay	Mortality (n)	Day 1 APS	Current Day APS	Change in APS	aROC
Day 1 Model	17% (17,440)	1.97	-	-	0.90
Day 2 Model	17% (14,034)	1.53	1.58	-	0.89
Day 3 Model	22% (8,860)	1.17	1.81	1.18	0.88
Day 4 Model	27% (5,884)	1.17	1.86	1.22	0.87
Day 5 Model	32% (4,164)	1.16	1.86	1.31	0.86
Day 6 Model	36% (3,137)	1.15	1.81	1.22	0.84
Day 7 Model	40% (2,489)	1.15	1.79	1.16	0.84

ICU: Intensive Care Unit
APS: APACHE III Acute Physiology Score
Change in APS: Previous day APS – Current day APS
n: Number of patients remaining in ICU
aROC: area under the receiver operating characteristic curve

The authors observed that although day one APS remained important in all models, the current day APS and the change in APS from the previous to the current day became consistently more important as the patient stayed longer in the ICU. Furthermore, it was reported that although discrimination remained high in all models, the area under the ROC curve consistently decreased as length of stay increased.

2.5.2 MPM

In order to improve the performance of risk prediction models on long-term stay patients, three different MPM models were developed (MPM_{24}, MPM_{48} and MPM_{OT}) using a 2,783 patient database collected in the adult general medical/surgical ICU at Baystate Medical Center.[50] The performance of these three models was compared directly to the performance of the MPM_0.

The MPM_{24}, MPM_{48} and MPM_{OT} all displayed good fit in the cohort of 948 patients whose length of stay was over 48 hours while the MPM_0, which was developed only on information available at admission, displayed poor calibration (H-L gof $p<0.0001$). A direct comparison of the MPM_{48} and MPM_{OT} suggested that MPM_{OT} was better at predicting conditional probabilities. In patients who ultimately died, MPM_{OT} was significantly better at predicting outcomes than MPM_{48} ($p=0.05$, McNemar's chi-square).

When the MPM models were updated in the MPM II paper, MPM II_{24} displayed good calibration in its developmental and validation data sets but calibration was poor when it was applied to cohorts of patients remaining in the ICU at 48 and 72 hours, (H-L gof $p<0.001$ for both 48 and 72 hour cohorts).[52] Development of the MPM II_{48} and MPM II_{72} was undertaken specifically to address the poor fit of the MPM II_{24} on longer stay patients.

Both MPM II_{48} and MPM II_{72} demonstrated good performance in long stay patients (MPM II_{48} H-L gof $p=0.31$, aROC=0.81 developmental, H-L gof $p=0.59$, aROC=0.796 validation and MPM II_{72} H-L gof $p=0.311$, aROC=0.79 developmental, H-L gof $p=0.408$, aROC=0.75 validation). Unfortunately, this paper did not compare the performance of MPM II_{24} with MPM II_{48} or MPM II_{72} on these long stay cohorts using a formal test of discrimination.

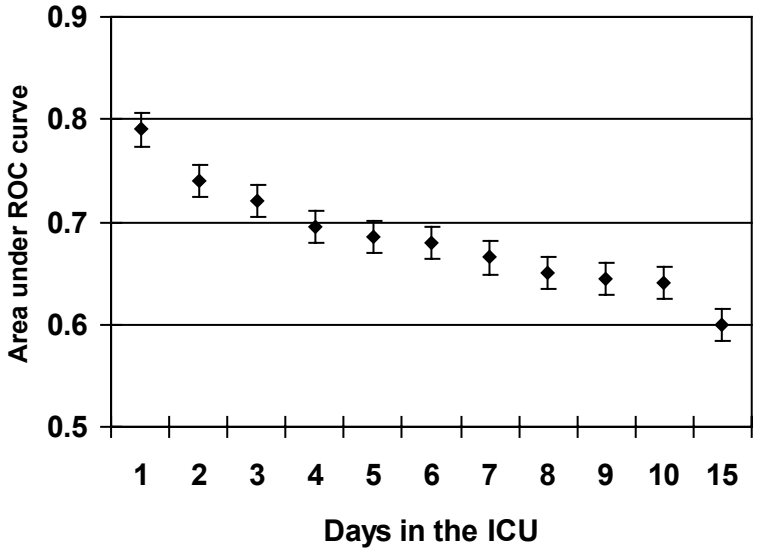
2.5.3 SAPS

Based on reports that both MPM and APACHE demonstrated decreased performance in long stay patients, a comprehensive independent evaluation of the SAPS scoring system was undertaken on a database composed of 8,059 patients collected from 24 centers throughout Europe.[59] In this study, SAPS was scored on each patient at admission using the original SAPS methodology[54] and a logistic regression model was developed using the approach outlined in the SAPS II paper.[55] To evaluate the accuracy of day one SAPS predictions in long stay patients, the area under the ROC curve was calculated for 10 different patient cohorts. The first cohort was composed of all eligible patients (N=8,059), the second was composed of all patients still remaining in the ICU on day two (N=5,992), the third was composed of all patients remaining in the ICU on day three (N=4,856) and so on for each of the initial 10 days of stay. The final cohort was composed of patients who remained in the ICU on day 15.

The SAPS predictive equation demonstrated both good calibration and discrimination (H-L gof $p=0.32$, aROC=0.79±0.01 developmental, H-L gof $p=0.53$, aROC=0.78±0.01 validation), however the area under curve systematically decreased in direct proportion to length of stay. The area under the ROC curve for the complete cohort of patients (0.79±0.01) was reported as being significantly higher than the area under the ROC curve for any other cohort of patients (Figure 1). This article demonstrated that the

predictive power of a SAPS model developed on day one data systematically decreased as length of stay increased. If a patient remained in the ICU for at least 15 days, outcome prediction based on day one SAPS was reported as being no better than a "toss of the coin".

Figure 1. Relationship between area under the ROC Curve and length of stay for the SAPS predictive equation.[59]



2.6 Neural Networks

Although logistic regression models have demonstrated acceptable predictive performance, some researchers believe that further attempts at refining existing systems will be fruitless and that understanding the complex patterns associated with mortality in the ICU may require the use of alternative mathematical approaches such as set theory or fuzzy logic.[12,13] Artificial neural networks (ANNs) are one such alternative mathematical approach that has generated a lot of interest in the field of medicine.

ANNs are pattern recognition algorithms that were originally developed by cognitive science researchers and are modeled after the biological structure of the human brain.[60] ANNs are widely used in the engineering disciplines of signals processing, image processing and control systems.[61] In the field of speech recognition, ANNs have consistently outperformed conventional regression techniques[15] and are capable of solving problems which even advanced artificial intelligence systems find difficult or

impossible to solve.[16] ANNs have even proven useful for gaining insights into the appropriate modeling of survival data.[62,63] A primer on ANN theory is presented in Appendix I.

In the field of medicine, image processing neural networks have been trained to diagnose hepatic masses[17] and breast tumors[19] with a level of accuracy similar to that of experienced clinicians. In 1993, an image processing neural network was trained to interpret ventilation-perfusion (V/Q) lung scans by exposing it to 100 consecutive V/Q scans using the subsequent pulmonary angiogram to confirm the diagnosis.[18] When presented with a series of 28 new V/Q scans, the network reportedly performed better than an experienced radiologist in the prediction of the chance of a pulmonary embolism.

A subsequent study combined V/Q scans and clinical assessment variables to compare the diagnostic performance of ANNs to experienced clinicians. The data for this study were abstracted from the 1,213 patient Prospective Investigation of Pulmonary Embolism Diagnosis (PIOPED) study.[64] The ANN was trained to diagnose pulmonary emboli on a 606 patient developmental database and performance was assessed using a 607 patient validation database. Based on the area under the ROC curves, the ANN performed as well as clinicians (area under the ROC curve: 0.83±0.013 vs. 0.85±0.017, p=NS) in the diagnosis of pulmonary emboli. The ability of ANNs to diagnose and predict outcomes has also been evaluated in various fields of medicine that do not rely heavily on diagnostic imaging.

An extensive assessment of the ability of ANNs to diagnose hepatic failure was undertaken using a database of 1,674 patients.[65] The data were divided evenly into developmental and validation databases containing 27 independent variables deemed to be predictive of liver failure. The outcome (liver failure) was diagnosed by specific laboratory tests that were not included among the 27 predictive variables.

A logistic regression model was developed using all variables with a univariate probability less than 0.05 and all possible 2-way interactions. The ANN included all 27 available input variables. Based on the area under the ROC curves for the validation database, there was no significant difference between the performance of the ANN and the logistic regression model (0.68±0.03 vs. 0.69±0.03, p=0.45) but the ANN did perform significantly better on the developmental database (0.80±0.03 vs. 0.74±0.03, p=0.04).

A second smaller study used a 144 patient database composed of alcoholic patients with severe liver disease to compare the ability of ANNs, clinical prediction rules, and logistic regression to predict in-hospital mortality.[66] Using a modified jackknife technique, the performance of the ANN was found to be superior to a widely accepted clinical prediction tool, the Maddrey score[67] (area under the ROC curve 0.81 vs. 0.74, p=0.04) but not significantly different from the logistic regression model (area under the ROC curve 0.82 vs. 0.78, p=0.3).

In cardiology, ANNs have been shown to be the most sensitive way to detect electrocardiographic arm lead reversal;[68] have been used to predict the risk of coronary artery disease;[69] to diagnose the onset of an acute myocardial infarction;[70,71] to predict hospital length of stay post-coronary care unit admission;[72] and to predict ICU length of stay after cardiac surgery.[73]

In a more recent study, the ability of an ANN to predict mortality post-cardiac surgery was compared with logistic regression.[74] This study used an extensive database of 4,782 patients who underwent coronary artery bypass surgery for model development. Two separate validation databases of 5,309 and 5,517 patients each were used to compare predictive performance. Both ANNs and logistic regression models were developed using 11 variables that had previously been shown to predict mortality post-cardiac surgery. These variables included; age, gender, ventricular ejection fraction, urgent/emergent surgery, previous cardiac surgery, presence of left main coronary artery disease, Canadian Cardiovascular Society angina class, recent myocardial infarction, diabetes, presence of chronic obstructive lung disease, and presence of peripheral vascular disease.

The ANN evaluated in this study was a back-propagation network developed with a learning rate of 0.1 and a momentum term of 0.1. The impact of using different activation functions (logistic, hyperbolic tangent etc.) and varying the number of hidden nodes was assessed on the predictive performance of the network.[75] The logistic regression model contained main-effects only. The performance of the ANN and the logistic regression model did not differ significantly when compared using the areas under the ROC curve for the validation database (0.78 vs. 0.77, p>0.10).

Accurate quality assessment programs are essential to the prevention, triage and treatment of severely injured trauma patients.[76] In one study based on data collected on

over 114,000 trauma patients, an ANN was shown to predict survival more accurately than the widely accepted Injury Severity Score (ISS). At a cut-off value of 0.5, the paper reports the ANN demonstrated higher sensitivity (0.99 vs. 0.91), higher specificity (0.50 vs. 0.45) and a higher overall accuracy (0.98 vs. 0.90) than the ISS.[77]

A second study used an 8,300 patient database composed of physiological and diagnostic criteria to assess the ability of an ANN to predict survival in trauma patients.[78] The back-propagation ANN was generated on a randomly selected 3,500 patient developmental database using the Revised Trauma Score (RTS), the ISS and age as input variables. The network was then applied to a 4,800 patient validation database and compared to the performance of the Trauma and Injury Severity Score (TRISS) and ASCOT. Both TRISS and ASCOT are highly validated logistic regression-based severity scores and use RTS, ISS and age as inputs. The ANN, TRISS and ASCOT all demonstrated poor fit (H-L gof $p<0.05$ all models) on the validation database. Using a cut-off value of 0.50, the sensitivity of the ANN, TRISS and ASCOTT was 0.904, 0.840 and 0.842 and the specificity was 0.972, 0.985 and 0.985 respectively. Based on a chi-square analysis, the ANN had both the highest sensitivity ($p<0.05$) and the lowest specificity ($p<0.05$).

In critical care medicine, two publications have objectively assessed the ability of ANNs to predict mortality in patients admitted to the ICU. In one of these studies, a genetic learning algorithm was used to optimize network architecture.[79] This study abstracted information from 258 ICU patients who were documented to have the Systemic Inflammatory Response Syndrome (SIRS). The database was randomly divided into a 168 patient developmental data set and a 90 patient validation data set. From 157 eligible candidate variables, a Classification and Regression Tree approach was used to select 11 variables for entry into the ANN and univariate analysis was used to select 9 variables for entry into the logistic regression model. Four significant 2-way interactions were added to the logistic regression model. Based on the aROC, the authors claimed the ANN outperformed the logistic regression model in the validation data set (0.863 vs. 0.753, no standard error or significance tests were reported).

In the publication that served as the pilot project for this thesis, a 420 patient database composed of the APS components of APACHE II collected on day 3 of ICU

stay was used to investigate the potential of an ANN to outperform a main-effects logistic regression model.[80] In this project, the back-propagation ANN exhibited discrimination superior to that of the multivariate logistic regression model in the 284 patient developmental data set (aROC 0.99 versus 0.92). The ANN performed with a positive predictive value of 0.98 and a negative predictive value of 1.0. When discrimination was compared on the 138 patient validation data set, the two methods performed equally well (aROC = 0.82).

Based on these results, the authors concluded that the neural network demonstrated the potential to outperform logistic regression with respect to the prediction of mortality in the ICU. It was also concluded that scoring on day three demonstrated the potential to improve performance over scoring on day one when predicting mortality of patients whose duration of stay was over 72 hours. It was proposed that using a neural network combined with day 3 data collection might result in risk estimates that could support clinical decisions and that these hypotheses should be tested in a project designed with sufficient power to assess them adequately. The copyright release for the reproduction of this pilot project is in Appendix II and the complete report of this pilot project is in Appendix III.

2.7 Research Questions

Logistic regression-based severity-of-illness scores are widely used to support quality assurance and resource utilization decisions. Although admission scores predict outcomes with acceptable accuracy when applied to the entire population of ICU patients, predictive performance decreases significantly as the patient's length of stay increases.

Artificial neural networks are a novel prediction tool that have previously been shown to outperform classical statistical techniques in certain situations. Based on a thorough review of the literature and the results of the pilot project, this thesis addresses the following primary research questions:

1) Can artificial neural networks perform mortality prediction in the ICU better than the currently used technique of multivariate logistic regression?

2) For patients with a duration of stay over 72 hours, will scoring on day 3 of ICU stay rather than day 1 increase the predictive performance of both artificial neural networks and logistic regression?

This project addresses the following secondary question:

1) Can artificial neural networks perform mortality prediction in the ICU significantly better than experienced clinicians, and thus demonstrate the potential to become a useful clinical decision support tool?

3. Methods

3.1 Location

This analytical observational study was undertaken in the Richard Ivey Critical Care Trauma Centre (CCTC), a 30 bed adult ICU that admits medical, surgical, neurological, cardiac surgery and trauma patients from the Victoria Campus of the London Health Sciences Centre (LHSC). The LHSC is a tertiary care teaching hospital associated with the University of Western Ontario.

3.2 Ethics

A two page overview outlining the purpose and methods of this project was submitted to the University of Western Ontario Ethics Review Board Office. The Chair of the Review Board For Health Sciences Research Involving Human Subjects determined that since this study was observational in nature and did not request or generate new information above or beyond what is normally accrued upon routine patient admission to the ICU, that a formal ethics review was unnecessary. The two page overview was approved on the conditions that the principal investigator (GSD) sign a confidentiality agreement with LHSC and that results would be published in such a way that individual patients could not be identified.

3.3 Patient Selection and Data Abstraction

Beginning March 1, 1994 and continuing through February 28, 1996, all patients admitted to the Richard Ivey CCTC were screened for entry into the study if they were present in the ICU at morning rounds three days after admission. If a patient were present at calendar day three morning rounds, and a discharge or withdrawal of care was not scheduled prior to the beginning of the rounds, they were formally entered into the study.

The pilot project for this thesis was conducted using a data set composed of APACHE II data elements.[80] APACHE III identified the 13 physiological variables of APACHE II plus five additional physiological variables, source of admission and age as independent predictors of mortality.[9] These components of the APACHE III scoring system were abstracted from patient charts following the methodology outlined by Knaus *et al.*[9] All data were collected on the morning of day three of stay. A complete list of

variable names is presented in Table 2. A reproduction of the data collection form can be found in Appendix IV.

Table 2 . Summary of variables collected, units used and abbreviations.

Demographic Variables	Units	Variable Name
Admission date - Date of birth	years	age
Gender	binary	male
Source of Admission:		
Non-operative admissions:		
Emergency room (referent)	binary	emerg
Hospital floor	binary	hosp
Transfer from another ICU	binary	icutrans
Transfer from another hospital	binary	hosptran
Operative admission:		
Emergent surgery	binary	or
Elective surgery	binary	elective
Physiologic Variables		
Glasgow Coma Scale	score from 3 to 15	gcs(s)(3)
Temperature	°C	temp(s)(3)
Systolic blood pressure	mmHg	sbp(s)(3)
Diastolic blood pressure	mmHg	dbp(s)(3)
Heart rate	beats per minute	hr(s)(3)
Respiratory rate	breaths per minute	rr(s)(3)
Urine output	L per day	uo(s)(3)
Fraction of inspired oxygen		FiO2(s)(3)
Partial pressure of arterial O_2	torr.	PaO2(s)(3)
Partial pressure of arterial CO_2	torr.	$PaCO_2$(s)(3)
pH		ph(s)(3)
Hemoglobin	g/L	hb(s)(3)
White blood cell count	$x10^{12}$/L	wbc(s)(3)
Platelets	$x10^{12}$/L	pts(s)(3)
Serum sodium	mmol/L	na(s)(3)
Serum potassium	mmol/L	k(s)(3)
Serum albumin	g/L	alb(s)(3)
Blood urea nitrogen	mmol/L	bun(s)(3)
Serum creatinine	Umol/L	creat(s)(3)
Glucose	mmol/L	glu(s)(3)
Bilirubin	Umol/L	bili(s)(3)

(s) suffix denotes 'Scaled' variable
(3) suffix denotes Day 3 physiologic data

For the purposes of this study, emergent surgery was defined as admission to the ICU immediately following unscheduled surgery for a life threatening condition. The CCTC is a level one trauma centre, so the majority of these admissions are as a result of trauma surgery. Other examples of procedures that would qualify for the designation of emergent used during this study include: surgery for a ruptured abdominal aortic aneurysm, surgery to drain an abdominal abscess, urgent cardiac bypass surgery or surgery for a ruptured cerebral aneurysm.

3.3.1 Consultant's Predicted Outcome

At the conclusion of day three morning rounds, the attending ICU consultant was requested to predict a patient's probability of surviving until ICU discharge, using an 11 point visual analog scale (See Appendix IV). A consultant predicted outcome was not requested on patients whose ICU discharge or withdrawal of care was scheduled during day three morning rounds.

3.4 Primary outcome of interest

The primary outcome of ICU-based mortality, as reflected by ICU discharge status, was collected by the CCTC's management information system which prospectively tracks status at discharge for all patients admitted to the ICU.

3.5 Database validation

The first stage of primary data integrity checking was undertaken by incorporating range restrictions into the data entry process, which was undertaken using a dBase© III [a] database. While the data were being entered into the database, a value recorded on the study code sheet as 'out of range' was entered as a missing value. More extensive validation with algorithms written to filter out biological impossibilities and obvious data entry transpositions was carried out using PC SAS® version 6.12[b]. Any values in conflict with the screening filters were re-entered directly from the study codebook.

If the study code book was found to have more than 4 missing or out of range values on any one particular patient, the patient's chart was requested from Medical Records and the missing information was re-abstracted and re-entered into the database.

Secondary validation was carried out by match-merging the study database with the ICU management information system (MIS) based on patient identification number (PIN) and date of admission. In cases where a merging match did not occur, patients were

---

[a] INPRISE Corporation, 100 Enterprise Way, Scotts Valley, CA 95066-3249, U.S.A.

[b] PC SAS® version 6.12, SAS Institute Inc., SAS Circle, PO Box 8000,Cary, NC, 27512-8000, U.S.A.

matched manually by comparing admission data, date of birth, and gender. This MIS database allowed validation of date of birth, ICU entry date, ICU discharge date and actual number of patients available for entry into the study. Since the MIS database is used for management and billing purposes, its entries are double verified and seldom incorrect. Conflicts with the study database were resolved by accepting the MIS database as correct.

3.6 Analysis

All statistical analysis was conducted using PC SAS Version 6.12 running under Microsoft Windows NT Workstation 4.0, Windows 95 or Windows 98 [a] on a Pentium II computer.

3.6.1 Descriptive statistics

Preliminary inspection of the data was undertaken using graphical and descriptive statistical techniques. Frequency plots and normal probability plots were generated to inspect data distributions. A formal test of association between the presence of missing values and the primary outcome of ICU-based mortality was undertaken. Although other approaches were considered, missing values were replaced with imputed average values.

Where the assumptions of normality were found to be inappropriate, descriptive data were presented as median and range. Otherwise, descriptive statistics were reported as frequencies or means and standard deviations.

3.6.2 Developmental and validation data sets

Using SAS, an index number was randomly generated from a distribution with mean of zero and a standard deviation of one for each patient record. The entire 1,149 patient database was sorted in ascending order by this randomly generated index number and the first 338 patients (30 percent) were removed and stored in a ‘validation’ database. The remaining 811 patient database was termed the developmental database and was used to develop both the logistic regression and neural network models. After all the models were developed, the performance of the models was compared based on their ability to predict outcomes in the validation data set.

---

[a] Microsoft Corporation, One Microsoft Way, Redmond, WA 98052-6399, U.S.A.

3.7 Logistic regression model development

Three distinct logistic regression models were developed. The first model considered demographic variables and day one physiologic variables for entry; the second model considered demographic variables and day three physiologic variables; and the third model was developed using demographic variables, day three physiologic variables plus the predicted output of the first (day one) model. All logistic regression modeling was performed using PROC Logist in PC SAS version 6.12[81] with the *events/count* statement in the model specification. The event of interest was ICU mortality coded as one for mortality and zero for survival to ICU discharge. The variable *count* was coded one for all cases. User selectable options were set to default values recommended by the software package authors.

3.7.1 Basic logistic regression modeling methodology

Step 1. Variable Selection

Univariate analysis was undertaken with each independent prognostic variable regressed against the primary outcome of ICU-based mortality. All variables with a univariate p-value less than 0.25 were considered for entry into the maximum model.[82]

Step 2. Specification of the Maximum Model

The maximum model contained all demographic and physiologic variables identified in *Step 1*. All possible two-way interactions between physiologic terms were eligible for entry into the maximum model.

Categorical demographic variables were represented as zero cell referent dummy variables. Since the pilot project revealed the potential for problems associated with multicollinearity, and since all previous investigations (APACHE, SAPs and MPM) have failed to document significant physiologic-demographic interaction terms, interaction terms between physiologic and demographic variables were not eligible for entry into the maximum model.

Step 3. Assessment for Multicollinearity

In the pilot project, standardization of the independent variables was required in order to avoid problems associated with multicollinearity. The maximum model was formally assessed for multicollinearity using Eigenanalysis with a condition index greater than or equal to 30 considered indicative of moderate to severe collinearity.[83]

Multicollinearity was addressed by standardizing all continuous independent variables,[84] the success of which was assessed using Eigenanalysis.

Step 4. Model Selection

After investigating the presence of multicollinearity, the maximum model was entered into a backwards selection model building process. In the first pass, all main effects were forced to stay in the model and the p-value to stay for interaction terms was set at 0.10.[82]

After all non-significant interaction terms were removed from the model, all main effects that were not associated with significant interaction terms were assessed for statistical significance using a back-ward elimination process and were left in the model if the likelihood ratio test p-value was less than 0.10. Main effects that contributed to significant interaction terms were always retained in the model regardless of the main effect's significance. All p-values used for decision making during the model selection process were based on likelihood ratio tests.

Step 5: Regression diagnostics

After generation of the final predictive model, performance on the developmental database was assessed. Calibration was assessed with the Hosmer-Lemeshow $\hat{C}$ goodness of fit statistic[85] and discrimination was reported using the c-statistic,[86] which is numerically equivalent to the area under the ROC curve.[87]

3.7.2 Day 1 logistic regression model development

The basic modeling approach described above was used to develop the day one logistic regression model. The variables that were considered for entry into the maximum model included demographic variables and day one physiologic variables (Table 2). This model is referred to as logistic model one (LM1).

3.7.3 Day 3 logistic regression model development

Two unique logistic regression models were developed using the data available by day three of stay. The first day three logistic model was developed using the basic modeling approach as described above and considered for eligibility all demographic variables plus day three physiologic variables (Table 2). This model is referred to as logistic model two (LM2).

A second day three model was developed as an extension to LM2. This model used the approach described in developing LM2 but it also incorporated the predicted output of LM1 as a main effect. To form its predictions this third logistic model (LM_{OT}) therefore incorporated all information that became available over time by day three of stay.

3.8 Artificial neural network model development

Four distinct ANNs were developed using NeuroShell 2, Release 3.0[a] software. In all cases, user selectable options were set to default values recommended by the software package authors.

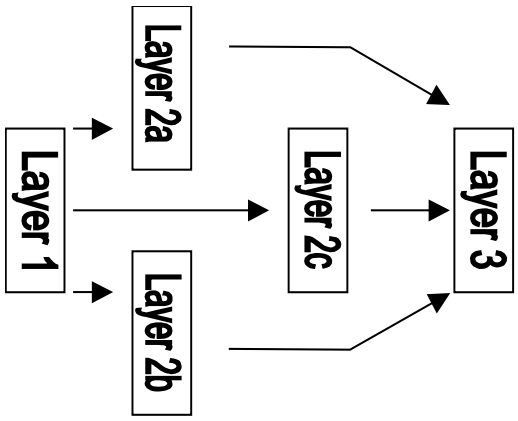
The four distinct ANNs were: 1) a back-propagation network developed on patient demographics and day one physiologic variables; 2) a back-propagation network developed on patient demographics and day three physiologic variables; 3) a back-propagation network developed on patient demographics, day one physiologic variables and day three physiologic variables; and 4) a genetic adaptive learning network developed on patient demographics and day three physiologic variables. All ANNs were developed on the same 811 patient developmental database used for the logistic regression model generation.

3.8.1 Basic back-propagation network development.

[a] Ward Systems Group, Inc., Executive Park West, 5 Hillcrest Dr., Frederick, MD. 21703, U.S.A.

The primary network topology used for this project was a three-layered back-propagation network with the hidden layer composed of 3 Slabs (Figure 2).

Figure 2. Back propagation neural network architecture.



Transfer Functions and Number of Nodes per Layer (n):

Layer 1:f(x)=x,n=27

Layer 2a*:f(x)=exp(x^2),n=14

Layer 2b*: f(x)=tanh(x),n=14

Layer 2c*: f(x)=1-exp(x^2),n=14

Layer 3:f(x)=1/(1+exp(-x)),n=1

Total Number of Hidden Nodes:

(27*14)+(27*14)+(27*14)+1=1135

*Layer 2 contains 3 slabs of 14 neurons each

All input variables were standardized (normalized) and then offered into the individual neurons of the input layer (Layer 1). The input layer contained one active transfer function (neuron) per input variable. The number of hidden transfer neurons (N_{hid}) in each of the three subsequent slabs was determined by Equation 1 where N_{input} is the number of input variables, N_{out} is the number of outcome variables, N_{cases} is the number of cases in the developmental data set and N_{slabs} is the number of slabs in the network under development. In all back-propagation networks considered here, N_{slabs} was fixed at the default of three and N_{out} was fixed at one.

Equation 1. Formula used to calculate the appropriate number of hidden nodes per network slab.[88]

$$N_{hid} = \frac{\frac{1}{2}(N_{input} + N_{out}) + \sqrt{N_{cases}}}{N_{slabs}}$$

Since the input variables were already standardized, Layer 1 used a simple linear transfer function to map information to Layer 2a, 2b and 2c. Layer 2a used a Gaussian transfer function to pass information to the output node, Layer 2b used the hyperbolic tangent function and Layer 2c used a Gaussian complement function to pass information to the output neuron. The output neuron (Layer 3) used a logistic function to map incoming information from Layer 2a, 2b and 2c into an estimate of the expected probability of mortality. Extensive evaluations of this approach are available elsewhere.[74] See Figure 2 for a complete listing of these transfer functions.

Learning rate refers to the minimum amount that each weight (regression parameter) may change after each training case, with larger learning rates indicating greater weight changes. The momentum term allows each weight change to be proportional to the magnitude of the previous change. Momentum allows the network to avoid local minimum in the error surface and speeds up convergence. The default parameters of a learning rate of 0.1, a momentum of 0.1 and initial node weights of 0.3 were used. Learning was terminated after 100,000 learning epochs had occurred without the generation of a new minimum average error on the test set. In the neural network literature, average network error is defined as the sum of the squares of the individual differences between the actual output values and the networks predicted output values. For further details on the back propagation algorithm, see Appendix I.

3.8.2 Artificial Neural Network 1: Day 1 model

The first back-propagation network was developed using all available demographic data plus day one physiological data. Since source of admission was offered

to the network as a dummy variable with Emergency Room coded as the referent, there were 28 unique input variables and thus 28 neurons in the input node.

3.8.3 Artificial Neural Network 2: Day 3 model

The second back-propagation network was developed using demographic data plus day three physiological data. Twenty-eight unique input variables were offered into this model.

3.8.4 Artificial Neural Network 3: Day 3 model over time

This model contained all data available by day three: demographic variables, day one physiologic variables and day three physiologic variables. This model therefore considered 48 unique input variables.

3.8.5 Artificial Neural Network 4: Genetic Learning Algorithm

The fourth neural network developed was a three layered general regression net that used a genetic adaptive learning algorithm (GenNet). The input data set was composed of all available demographic variables (age, gender and source of admission) and the twenty physiologic variables collected on day three of stay. The network was trained on the 811 pattern training set and calibrated on the 338 patient validation set.

The transfer function in the input layer of GenNet was a one to one mapping of the standardized input variables to the neurons in the second layer. When using the genetic adaptive learning algorithm, the number of neurons in the second layer is set to the sum of the number of available training and validation cases. The number of hidden neurons was therefore set to 1,149 (811 training cases+338 validation cases=1,149 hidden neurons). The default smoothing factor of 0.2468 was used for all connections and the genetic breeding pool size was set to 100. Development was terminated when average validation network error did not improve more than one percent after a 20 generation learning cycle. For further reading on the genetic adaptive learning algorithm, see Appendix V.

3.9 Sample Size Considerations

Since this study was driven by the application of neural network theory, it was decided that our sample size calculations should be determined by concerns arising from that field.

Based on information generated from the pilot project, it was determined that a reasonable number of inputs to the neural network would be 44 variables (at most 36 APS measurements plus the 8 demographic measurements of the APACHE III score). Since the pilot study required 18 hidden nodes, it was a reasonable assumption that the future study network would require approximately 18 hidden nodes. Given a network of 44 inputs, 18 hidden nodes and 1 output node, there would be (44+1)*18 = 810 interconnections or modifiable weights.

Although there are no formal methods to determine the size of a training set required for a network to achieve acceptable performance, many researchers have found that it begins to appear when the net is presented with *at least* one training event per modifiable weight.[89] With 810 interconnections or modifiable weights the expected network would require a training set of at least 810 events to achieve acceptable performance. Since the training (developmental) data set would be composed of two-thirds of the total data set, a total of at least 1,215 eligible patients was estimated to be required in the complete database.

3.10 Feasibility

3.10.1 Subject Availability

During the 6 month sampling period for the pilot project, 614 patients were admitted to the CCTC. Of the 614 admissions, 422 had a duration of stay over 72 hours. This translates into 70 eligible patients per month. Since a minimum sample size of 1,215 patients is required, then in 18 months of sampling, 1,260 patients would become eligible for the study. This would be the minimum acceptable collection period to expect reasonable performance from the neural network.

3.11 Comparing Predictive Performance: The Validation Database

Primary comparisons were conducted based on performance on the 338 patient validation database. Goodness of fit for all models was assessed using the H-L $\hat{C}$ statistic with 8 degrees of freedom on a chi-square distribution in the developmental database and 10 degrees of freedom in the validation database.[82,85] Discrimination was assessed using the area under the ROC curve. The area under the ROC curve, along with

its associated standard error, was calculated via the Dorfman-Berbaum-Metz maximum likelihood method [90] using the computer program ROCKIT [a].

Primary comparisons of the areas under the ROC curves were conducted by using a formal statistical test developed for correlated (paired) data.[91]

[a] ROCKIT 0.9B – beta version (March 1998), Charles E. Metz, Department of Radiology and the Franklin McLean Memorial Research Institute, The University of Chicago, Chicago, Ill, 60637.

4. Results

4.1 Patient Selection

Of the 3,728 patients admitted to the CCTC during the 24 month period from March 1, 1994 through February 28, 1996, 1,263 patients were present in the ICU at 48 hours and were therefore screened for study eligibility. Eighty-two of these 1,263 patients had an ICU discharge scheduled prior to data collection on the morning of day three. The remaining 1,181 patients had an 'expected length of stay greater than 72 hours' at the time of data collection and were therefore formally entered into the study.

Information was abstracted prospectively from patient records and recorded in the study logbook. A complete listing of all variables collected, along with their abbreviations, is presented in Table 2.

4.2 Database Validation

Initial data validation was performed by simple range checks at the time of input into the dBase III database. Extensive validity checking in SAS revealed that, at the time of original data collection, approximately 300 of these 1,181 patient records had more than four missing or out of range values. Patient identification numbers for these 300 patients were submitted to medical records and a retrospective chart audit was performed to fill in the missing data.

Of the 300 charts requested, 32 charts could either not be found by medical records or were found to be significantly incomplete at the time of retrospective abstraction. A chart was considered significantly incomplete when physiologic values for at least an entire day were missing. Due to the large amount of missing data for each of these 32 patients, the only variables consistently available for comparison were the outcomes recorded by the CCTC database. The median length of stay for these 32 patients was 3.9 days, (range 3 to 73 days) and the ICU-based mortality was 28 percent. These values were not significantly different from those of the 1,149 patients on whom complete information was available (median length of stay 6 days with a range of 3 to 179 days, ICU based mortality 14.4 percent). Due to the unavailability of significant portions of information on these 32 patients, they were excluded from further analysis.

The remaining 1,149 patients comprised the final data set. A directed chart audit revealed that the missing value rate for the entire data set was 6.75 percent. On further investigation two variables, bilirubin collected on day one and bilirubin collected on day three, were responsible for 60 percent of all missing values. Since clinicians from the CCTC do not routinely order these tests on all patients, these two variables were excluded from further analysis, reducing the overall missing value rate to 2.62 percent.

An evaluation of the relationship between missing values and outcome was undertaken. Using Fisher's Exact test and controlling for multiple comparisons using a Bonferonni correction revealed that missing values for three different variables were associated with an improved outcome. Having a missing value for blood glucose on day one was significantly associated with a decreased risk of mortality (2.45% vs. 16.07%, relative risk = 0.15, p=1.67E-07), as was a missing blood glucose value on day three (3.75% vs. 15.82%, relative risk = 0.24, p=7.56E-6). A missing blood urea nitrogen (BUN) value on day three (3.24% vs. 16.23%, relative risk = 0.20, p=2.25E-07) was also found to be associated with an improved outcome. For these three and for all other variables, the mean value was associated with improved outcomes. All missing values were replaced with calculated mean values.

4.2.1 Descriptive statistics

The overall mortality rate for the 1,149 patients in the final database was 14.4 percent. Descriptive statistics for all demographic variables are presented in Table 3. Descriptive statistics for all physiologic variables are presented in Table 4.

Table 3. Descriptive statistics on demographic variables in the complete 1,149 patient database.

Variable name	Results
Age	61.8±17.25 years
Gender	38.6% females
Length of stay	6 (3 to 179) days
Non-operative admissions	
Emergency room	292 (25.4%)
Hospital floor	238 (20.7%)
Transfer from another hospital	65 (5.7%)
Transfer from another ICU	65 (5.7%)
Post-operative admissions	
Elective surgery	252 (21.9%)
Emergent surgery	237 (20.6%)
Total patients	1,149

Length of stay is presented as median (range).

Table 4. Descriptive statistics on physiologic variables in the complete 1,149 patient database.

Variables Collected on Day 1 Name	n	Mean	Standard Deviation	Minimum	Maximum
Glasgow coma scale	1149	7.38	4.47	3	15
Albumin	1149	27.55	6.98	2	48
Bilirubin	57	14.45	10.16	3	49
Blood urea nitrogen	1149	9.73	7.05	0.57	87
Creatinine	1149	139.72	117.54	10	1470
Diastolic blood pressure	1149	62.71	22.81	8	150
FiO2	1149	0.75	0.25	0.21	1
Glucose	1149	15.62	13.64	0.50	309
Hemoglobin	1149	108.83	26.05	9.9	201
Heart rate	1149	102.83	32.05	17	300
Potassium	1149	3.91	0.75	1.8	7.5
Sodium	1149	138.09	6.00	112	216
PaCO2	1149	36.84	12.71	10	311
PaO2	1149	214.40	140.47	10	652
pH	1149	7.39	0.09	6.95	7.69
Platelets	1149	209.36	108.82	1	678
Respiratory rate	1149	16.19	8.28	3	56
Systolic blood pressure	1149	132.50	44.97	15	293
Temperature	1149	36.62	1.74	29.7	42.6
Urine output	1149	1471.38	931.70	2	7801
White blood cell count	1149	14.38	15.73	0	195
Variables Collected on Day 3 Name	**n**	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
Glasgow Coma Scale	1149	11.51	3.75	3	15
Albumin	1149	28.81	6.05	8	46
Bilirubin	41	24.68	16.94	3	70
Blood urea nitrogen	1149	11.88	8.52	0.53	53
Creatinine	1149	132.68	130.43	10	1980
Diastolic blood pressure	1149	63.19	18.14	6	150
FiO2	1149	0.47	0.16	0.21	1
Glucose	1149	13.29	12.12	0.5	303
Hemoglobin	1149	97.84	20.49	9	177
Heart rate	1149	102.08	29.56	4	283
Potassium	1149	4.01	0.58	1.65	8.8
Sodium	1149	138.17	5.32	113	160
PaCO2	1149	41.10	14.31	10	377
PaO2	1149	95.67	37.73	10	460
pH	1149	7.41	0.07	6.97	7.98
Platelets	1149	168.53	95.23	1	663
White blood cell count	1149	13.59	10.37	0.6	152
Respiratory rate	1149	19.15	7.30	3	60
Systolic blood pressure	1149	138.88	36.35	7	246
Temperature	1149	37.56	1.01	30.6	41
Urine output	1149	1413.97	818.93	0	5800

All variables were recorded using standard SI units except blood pressure, which was recorded in mmHg. See Table 2 for a complete listing.

4.3 Logistic regression model development

4.3.1 LM1: Logistic Model based on day one data

Step 1. Variable Selection

Table 5 presents the results of the descriptive analysis of the demographic variables. Source of admission was the only demographic variable that was associated with outcome ($p<0.25$) and therefore qualified for entry into the maximum model.

Table 5. Results of univariate logistic regression analysis of the association between demographic factors and outcome in the 811 patient developmental database.

Variable Name	Parameter Estimate	Standard Error	Wald Chi-Square	Probability value	Odds Ratio
Age	0.0012	0.0049	0.0608	0.8052	1.001
Gender	-0.1297	0.1712	0.5736	0.4488	0.878
Source of Admission					
Emergency room	(referent)				
Hospital Floor	0.8344	0.2411			2.303
Another Hospital	0.0967	0.4204			1.102
Another ICU	0.6740	0.3610			1.962
Elective Surgery	-0.6309	0.3176			0.532
Emergent Surgery	0.4958	0.2523			1.642
Likelihood Ratio Chi-square			**33.550 with 5 DF (p=0.0001)**		

The following day one physiologic variables were eligible ($p<0.25$) for entry into the maximum model: albumin (alb), blood urea nitrogen (bun), creatinine (creat), diastolic blood pressure (dbp), fraction of inspired oxygen (FiO2), glucose (glu), heart rate (hr), partial pressure of arterial oxygen (PaO2), pH, platelets (pts), respiratory rate (rr), urinary output (uo), and white blood cell (wbc) count. A detailed listing of univariate regression results for day one physiologic variables is presented in Table 6.

Table 6. Results of univariate logistic regression analysis of the association between physiologic variables collected on Day 1 and outcome in the 811 patient developmental database.

Variable Name	Parameter Estimate	Standard Error	Wald Chi-Square	Probability Value	Odds Ratio
Glasgow coma scale	0.0021	0.0235	0.0086	0.9263	1.002
Albumin	**-0.0385**	**0.0149**	**6.7206**	**0.0095**	**0.962**
Blood urea nitrogen	**0.0749**	**0.0134**	**30.9711**	**0.0001**	**1.078**
Creatinine	**0.0019**	**0.0006**	**8.5880**	**0.0034**	**1.002**
Diastolic blood pressure	**-0.0062**	**0.0048**	**1.6657**	**0.1968**	**0.994**
FiO2	**0.5620**	**0.4276**	**1.7274**	**0.1887**	**1.754**
Glucose	**-0.0235**	**0.0121**	**3.7955**	**0.0514**	**0.977**
Hemoglobin	-0.0008	0.0039	0.0518	0.8199	0.999
Heart rate	**0.0074**	**0.0032**	**5.1009**	**0.0239**	**1.007**
Potassium	0.1016	0.1380	0.5419	0.4617	1.107
Sodium	-0.0094	0.0184	0.2654	0.6064	0.991
PaCO2	0.0083	0.0103	0.6541	0.4186	1.008
PaO2	**-0.0028**	**0.0008**	**10.9845**	**0.0009**	**0.997**
PH	**-3.7569**	**1.0583**	**12.6028**	**0.0004**	**0.023**
Platelets	**0.0032**	**0.0009**	**11.1473**	**0.0008**	**1.003**
Respiratory rate	**0.0312**	**0.0116**	**7.2152**	**0.0072**	**1.032**
Systolic blood pressure	-0.0023	0.0023	1.0151	0.3137	0.998
Temperature	0.0502	0.0612	0.6727	0.4121	1.051
Urine output	**0.0002**	**0.0001**	**7.6254**	**0.0058**	**1.000**
White blood cell count	**0.0086**	**0.0048**	**3.1965**	**0.0738**	**1.009**

Bold indicates variables with probability values less than 0.25.

Step 2. Specification of the Maximum Model

The maximum model contained the following variables as main effects: source of admission, a dummy variable with emergency room as referent, alb, bun, creat, dbp, FiO2, glu, hr, PaO2, pH, pts, rr, uo and wbc. The maximum model also contained all possible 2-way interactions between the physiologic variables. A complete list of interaction terms generated for this maximum model can be found in Appendix VI.

Step 3. Assessment for Multicollinearity

Eigenanalysis of the maximum model revealed a condition index of 395, which indicated the presence of severe multicollinearity. The continuous variables in the maximum model were standardized and 2-way interactions were recalculated. After standardization, the condition index was 6.87, indicating that standardization had addressed the issues of multicollinearity.

Step 4. Model Selection

After all non-significant interaction terms were removed from the model, main effects that were not associated with significant interaction terms were tested for significance using likelihood ratio tests. No main effect terms were removed. Table 7 shows the final results of the modeling process.

Table 7. Final regression parameters and associated probability values for multivariate Logistic Model 1 (LM1).

Variable Name	Parameter Estimate	Standard Error	Wald Chi-Square	Probability Value	Odds Ratio
INTERCPT	-3.1889	0.3193	99.7388	0.0001	
Hospital Floor	1.1986	0.3725	10.3520	0.0013	3.316
Hospital Transfer	-0.3791	0.6381	0.3531	0.5524	0.684
ICU Transfer	0.7851	0.5618	1.9530	0.1623	2.193
Elective Surgery	0.0842	0.4146	0.0412	0.8391	1.088
Emergent Surgery	1.1459	0.3770	9.2371	0.0024	3.145
Albumin(s)	-0.2155	0.1272	2.8702	0.0902	0.806
Blood urea nitrogen(s)	0.7108	0.1239	32.9162	0.0001	2.036
Diastolic blood pressure(s)	-0.0658	0.1334	0.2434	0.6217	0.936
FiO2(s)	0.3473	0.1431	5.8875	0.0152	1.415
Glucose(s)	-0.8760	0.2614	11.2274	0.0008	0.416
Heart rate(s)	0.0209	0.1363	0.0235	0.8782	1.021
PaO2(s)	-0.3812	0.1565	5.9306	0.0149	0.683
pH(s)	-0.3487	0.1305	7.1421	0.0075	0.706
Platelets(s)	0.4616	0.1296	12.6778	0.0004	1.587
Respiratory rate(s)	0.1981	0.1413	1.9660	0.1609	1.219
Urine output(s)	0.3752	0.1135	10.9229	0.0009	1.455
White blood cells(s)	0.0840	0.1354	0.3854	0.5347	1.088
Albumin(s)*diastolic blood Pressure(s)	0.3830	0.1240	9.5349	0.0020	1.467
Albumin(s)*white blood cells(s)	-0.3200	0.1557	4.2272	0.0398	0.726
Blood urea nitrogen(s)*glucose(s)	0.7321	0.2159	11.4961	0.0007	2.080
Diastolic blood pressure(s) *heart rate(s)	-0.3258	0.1356	5.7701	0.0163	0.722
FiO2(s)*urine output(s)	-0.3673	0.1224	9.0122	0.0027	0.693
Heart rate(s)*platelets(s)	0.2425	0.1378	3.0971	0.0784	1.274
PaO2(s)*pH(s)	0.3828	0.1341	8.1530	0.0043	1.466
PaO2(s)*respiratory rate(s)	0.2897	0.1550	3.4960	0.0615	1.336
pH(s)*respiratory rate(s)	0.4623	0.1399	10.9249	0.0009	1.588
Platelets(s)*respiratory rate(s)	-0.3120	0.1260	6.1291	0.0133	0.732

For a complete list of variable names, see Table 2.
(s) denotes standardized variable.
Variable_name1*variable_name2 denotes interaction term.

Step 5. Regression Diagnostics

On the 811 patient development data set, the final day one logistic model revealed a c-statistic of 0.850 and Hosmer and Lemeshow $\hat{C}$ chi-square of 7.0324, which has an associated p-value of 0.5331 with eight degrees of freedom.

4.3.2 LM2: Logistic Model based on day three data

Step 1. Variable Selection

Source of admission was the only demographic variable that was associated with outcome ($p<0.25$) and qualified for entry into the maximum model.

The following physiological variables collected on day three of stay were associated with mortality with a p-value less than 0.25: day 3 Glasgow Coma Scale score (gcs3), day 3 blood urea nitrogen (bun3), day 3 creatinine (creat3), day 3 diastolic blood

pressure (dbp3), day 3 fraction of inspired oxygen (FiO23), day 3 heart rate (hr3), day 3 sodium (na3), day 3 partial arterial pressure of carbon dioxide (PaCO23), day 3 partial arterial pressure of oxygen (PaO23), day 3 pH (pH3), day 3 platelet count (pts3), day 3 respiratory rate (rr3), day 3 temperature (temp3), day 3 urinary output (uo3), and day 3 white blood cell count (wbc3) (Table 8).

Table 8. Results of univariate logistic regression analysis of the association between physiologic variables collected on Day 3 and outcome in the 811 patient developmental database.

Variable Name	Parameter Estimate	Standard Error	Wald Chi-Square	Probability Value	Odds Ratio
Glasgow coma scale	**-0.2452**	**0.0277**	**78.5541**	**0.0001**	**0.783**
Albumin	**-0.0328**	**0.0175**	**3.5025**	**0.0613**	**0.968**
Blood urea nitrogen	**0.0335**	**0.0106**	**9.9508**	**0.0016**	**1.034**
Creatinine	**0.0044**	**0.0007**	**39.8221**	**0.0001**	**1.004**
Diastolic blood pressure	**-0.0272**	**0.0065**	**17.2520**	**0.0001**	**0.973**
FiO2	**3.4357**	**0.5325**	**41.6225**	**0.0001**	**31.054**
Glucose	-0.0091	0.0128	0.5086	0.4757	0.991
Hemoglobin	0.0052	0.0050	1.0727	0.3003	1.005
Heart rate	**0.0103**	**0.0035**	**8.2809**	**0.0040**	**1.010**
Potassium	-0.1836	0.1908	0.9264	0.3358	0.832
Sodium	**0.0796**	**0.0193**	**17.0633**	**0.0001**	**1.083**
PaCO2	**-0.0358**	**0.0117**	**9.4100**	**0.0022**	**0.965**
PaO2	**-0.0043**	**0.0031**	**1.9174**	**0.1661**	**0.996**
pH	**-2.1791**	**1.3005**	**2.8075**	**0.0938**	**0.113**
Platelets	**0.0028**	**0.0011**	**6.1650**	**0.0130**	**1.003**
Respiratory rate	**-0.0476**	**0.0159**	**8.9827**	**0.0027**	**0.953**
Systolic blood pressure	-0.0027	0.0029	0.8335	0.3613	0.997
Temperature	**-0.2468**	**0.0982**	**6.3120**	**0.0120**	**0.781**
Urine output	**-0.0003**	**0.0001**	**5.2272**	**0.0222**	**1.000**
White blood cell count	**0.0154**	**0.0078**	**3.8439**	**0.0499**	**1.015**

Bold denotes probability value less than 0.25.

Step 2. Specification of the Maximum Model

The maximum model for LM2 contained the following main effects: source of admission, gcs3, bun3, creat3, dbp3, FiO23, hr3, na3, PaCO23, PaO23, pH3, pts3, rr3, temp3, uo3 and wbc3. All possible two-way interactions between the physiologic variables were also offered to the maximum model (Appendix VII).

Step 3. Assessment for Multicollinearity

The maximum model, as described above, was assessed for multicollinearity using Eigenanalysis and a condition index of 585.12 revealed the presence of severe collinearity. Continuous variables were standardized and all two-way interactions were recalculated using the standardized main effects. Eigenanalysis of this standardized maximum model revealed a condition index of 12.35, indicating that standardization had addressed the presence of multicollinearity.

Step 4. Model Selection

After all non-significant interaction terms were removed from the model using a backwards elimination process, all main effects that were not associated with significant interaction terms were tested for significance using a likelihood ratio test. No main effects were removed. The final results of the backwards selection process are presented in Table 9.

Table 9. Final regression parameters and associated probability values from multivariate Logistic Model 2 (LM2).

Variable Name	Parameter Estimate	Standard Error	Wald Chi-Square	Probability Value	Odds Ratio
INTERCPT	-4.106	0.427	92.423	0.0001	.
Hospital floor	1.153	0.451	6.541	0.0105	3.170
Hospital transfer	0.392	0.743	0.278	0.5978	1.480
ICU transfer	0.931	0.661	1.984	0.1590	2.538
Elective surgery	0.451	0.529	0.726	0.3941	1.570
Emergent surgery	1.206	0.454	7.055	0.0079	3.342
Glasgow coma scale(s3)	-1.160	0.189	37.427	0.0001	0.313
Albumin(s3)	-0.067	0.180	0.140	0.7074	0.935
Blood urea nitrogen(s3)	0.052	0.191	0.076	0.7824	1.054
Creatinine(s3)	0.219	0.207	1.126	0.2886	1.246
Diastolic blood pressure(s3)	-0.726	0.182	15.831	0.0001	0.483
FiO2(s3)	0.501	0.163	9.444	0.0021	1.652
Heart rate(s3)	0.301	0.185	2.660	0.1029	1.352
Sodium(s3)	0.454	0.165	7.577	0.0059	1.575
PaCO2(s3)	-0.665	0.274	5.868	0.0154	0.514
PaO2(s3)	-0.529	0.229	5.334	0.0209	0.589
PH(s3)	0.043	0.158	0.075	0.7841	1.044
Platelets(s3)	0.163	0.180	0.822	0.3644	1.178
Respiratory rates(s3)	-0.519	0.181	8.207	0.0042	0.595
Temperature(s3)	-0.536	0.185	8.423	0.0037	0.585
Urine output(s3)	-0.126	0.172	0.537	0.4635	0.881
White blood cells(s3)	0.214	0.197	1.182	0.2769	1.240
Glasgow coma scale(s3)*albumin(s3)	-0.381	0.169	5.076	0.0243	0.683
Glasgow coma scale(s3)*blood urea(s3)	0.359	0.214	2.805	0.0939	1.432
Glasgow coma scale(s3)*creatinine(s3)	-0.544	0.212	6.598	0.0102	0.580
Glasgow coma scale*platelets(s3)	-0.434	0.176	6.033	0.0140	0.648
Albumin(s3)*diastolic pressure(s3)	0.388	0.176	4.867	0.0274	1.475
Blood urea nitrogen(s3)*FiO2(s3)	0.287	0.158	3.309	0.0689	1.333
Blood urea nitrogen(s3) *heart rate(s3)	0.526	0.200	6.911	0.0086	1.692
Blood urea nitrogen(s3) *respiratory rate(s3)	0.364	0.175	4.317	0.0377	1.440
Creatinine(s3)*sodium(s3)	-0.285	0.121	5.527	0.0187	0.752
Creatinine(s3)*PaO2(s3)	0.572	0.229	6.209	0.0127	1.773
Creatinine(s3)*temperature(s3)	-0.581	0.209	7.687	0.0056	0.559
Creatinine(s3)*urine output(s3)	-0.391	0.184	4.470	0.0345	0.676
Diastolic pressure(s3) *white blood cells(s3)	0.511	0.209	5.992	0.0144	1.668
FiO2(s3)*pH(s3)	0.280	0.139	4.081	0.0434	1.324
FiO2(s3)*temperature(s3)	0.364	0.133	7.475	0.0063	1.439
FiO2(s3)*urine output(s3)	0.383	0.140	7.476	0.0063	1.467
FiO2(s3)*white blood cells(s3)	-0.502	0.215	5.424	0.0199	0.605
Heart rate(s3)*respiratory rate(s3)	-0.393	0.143	7.462	0.0063	0.675
PacO2(s3)*paO2(s3)	-0.567	0.258	4.827	0.0280	0.567
PacO2(s3)*temperature(s3)	-0.534	0.209	6.516	0.0107	0.586
PacO2(s3)*white blood cells(s3)	0.423	0.218	3.741	0.0531	1.527
PaO2(s3)*pH(s3)	-0.557	0.143	15.156	0.0001	0.573
PaO2(s3)*white blood cells(s3)	-0.792	0.220	12.955	0.0003	0.453
PH(s3)*temperature(s3)	0.251	0.134	3.499	0.0614	1.286
Platelets(s3)*urine output(s3)	0.680	0.167	16.595	0.0001	1.975
Temperature(s3)*urine output(s3)	0.555	0.154	12.944	0.0003	1.743

(s3) denotes standardized Day 3 variable. Variable_name1*variable_name2 denotes interaction term.

Step 5. Regression Diagnostics

Based on performance on the developmental database, the c-statistic for LM2 was 0.927. The H-L goodness-of-fit $\hat{C}$ statistic was 10.059 with 8 df (p=0.2609), which indicated good fit.

4.3.3 LM_{OT}: Logistic Model constructed over time

Step 1. Variable Selection

Source of admission was the only demographic variable that was associated with outcome ($p<0.25$) and qualified for entry into the maximum model.

The following physiological variables collected on day three of stay were associated with mortality with a p-value less than 0.25: Glasgow Coma Scale score (gcs3), blood urea nitrogen (bun3), creatinine (creat3), diastolic blood pressure (dbp3), fraction of inspired oxygen (FiO23), heart rate (hr3), sodium (na3), partial arterial pressure of carbon dioxide (PaCO23), partial arterial pressure of oxygen (PaO23), pH (pH3), platelet count (pts3), respiratory rate (rr3), temperature (temp3), urinary output (uo3), and white blood cell count (wbc3).

Step 2. Specification of the Maximum Model

The maximum model for LM_{OT} contained the following main effects: pred(LM1), source of admission, gcs3, bun3, creat3, dbp3, FiO23, hr3, na3, PaCO23, PaO23, pH3, pts3, rr3, temp3, uo3 and wbc3. All possible two-way interactions between the physiologic variables were offered to the maximum model. A complete listing of the interaction terms generated for LM_{OT} are listed in Appendix VII.

Step 3. Assessment for Multicollinearity

The maximum model described above was assessed for multicollinearity using Eigenanalysis. The initial condition index of LM_{OT} was 585.19, which indicated the presence of severe collinearity. All continuous physiological variables were standardized and 2-way interactions were recalculated. Eigenanalysis demonstrated the standardized maximum model had a condition index of 12.36, indicating that the problems associated with collinearity had been addressed.

Step 4. Model Selection

After the initial backwards elimination pass, main effects not associated with significant interaction terms were inspected for significance. Based on a non-significant likelihood ratio test, Source of Admission [(-2 log likelihood for model with Source) 354.369 – (-2 log likelihood for model without Source) 347.564 = 6.805 with 5 df, p=0.2356] was removed from the model. With removal of Source of Admission, INT3s73 (FiO2s3*temps3) became non-significant (p=0.1167) and was also removed based on the results of a likelihood ratio test. Regression coefficients for the final LM_{OT} model are in Table 10.

Step 5. Regression Diagnostics

Based on the 811 patient developmental database, the c-statistic for LM_{OT} was 0.947 and the H-L goodness-of-fit $\hat{C}$ statistic was 4.2021 with 8 df (p=0.8384).

Table 10. Final regression parameters and associated probability values from multivariate Logistic Model Over Time (LM_{OT}).

Variable Name	Parameter Estimate	Standard Error	Wald Chi-Square	Probability Values	Odds Ratio
INTERCEPT	-3.7003	0.3201	133.5989	0.0001	.
Predicted outcome from LM1	1.3931	0.1850	56.7243	0.0001	4.027
Glasgow coma scale(s3)	-0.7202	0.2008	12.8704	0.0003	0.487
Albumin(s3)	-0.0794	0.1927	0.1697	0.6804	0.924
Blood urea nitrogen(s3)	-0.2368	0.1991	1.4153	0.2342	0.789
Creatinine(s3)	0.4292	0.1952	4.8344	0.0279	1.536
Diastolic blood pressure(s3)	-0.6271	0.1901	10.8854	0.0010	0.534
FiO2(s3)	0.5581	0.1402	15.8481	0.0001	1.747
Heart rate(s3)	0.3570	0.1937	3.3975	0.0653	1.429
Sodium(s3)	0.3010	0.1903	2.5016	0.1137	1.351
PaCO2(s3)	-0.1416	0.2278	0.3866	0.5341	0.868
PaO2(s3)	-0.5861	0.2555	5.2622	0.0218	0.556
pH(s)	0.5100	0.2094	5.9332	0.0149	1.665
Platelets(s3)	-0.1460	0.2050	0.5075	0.4762	0.864
Respiratory rate(s3)	-0.5471	0.1942	7.9344	0.0049	0.579
Temperature(s3)	-0.3964	0.1815	4.7737	0.0289	0.673
Urine output(s3)	-0.3190	0.2035	2.4571	0.1170	0.727
White blood cells(s3)	0.0369	0.1750	0.0445	0.8329	1.038
Glasgow coma scale(s3)*albumin(s3)	-0.3929	0.1882	4.3579	0.0368	0.675
Glasgow coma scale(s3) *blood urea nitrogen(s3)	0.5222	0.2290	5.1977	0.0226	1.686
Glasgow coma scale(s3) *creatinine(s3)	-0.3945	0.2127	3.4408	0.0636	0.674
Glasgow coma scale(s3)*sodium(s3)	-0.3549	0.1512	5.5087	0.0189	0.701
Glasgow coma scale(s3) *platelets(s3)	-0.5241	0.2121	6.1087	0.0135	0.592
Blood urea nitrogen(s3) *heart rate(s3)	0.6146	0.2087	8.6759	0.0032	1.849
Creatinine(s3)*sodium(s3)	-0.3195	0.1231	6.7365	0.0094	0.727
Creatinine(s3)*PaO2(s3)	0.3906	0.2335	2.7977	0.0944	1.478
Diastolic blood pressure(s3) *pH(s3)	0.3709	0.1546	5.7557	0.0164	1.449
Diastolic blood pressure(s3) *temperature(s3)	-0.4048	0.1771	5.2241	0.0223	0.667
Diastolic blood pressure(s3) *white blood cells(s3)	0.6817	0.2816	5.8602	0.0155	1.977
FiO2(s3)*PaCO2(s3)	-0.4108	0.1561	6.9289	0.0085	0.663
FiO2(s3)*urine output(s3)	0.3114	0.1328	5.4938	0.0191	1.365
PaCO2*PaO2(s3)	-0.5122	0.2641	3.7602	0.0525	0.599
PaO2(s3)*pH(s3)	-0.3646	0.1653	4.8617	0.0275	0.694
PaO2(s3)*white blood cells(s3)	-0.7966	0.2565	9.6494	0.0019	0.451
PH(s3)*respiratory rate(s3)	-0.3038	0.1637	3.4454	0.0634	0.738
PH(s3)*temperature(s3)	0.2337	0.1080	4.6798	0.0305	1.263
Platelets(s3)*urine output(s3)	0.7969	0.2154	13.6869	0.0002	2.219
Temperature(s3)*urine output(s3)	0.3562	0.1674	4.5297	0.0333	1.428
Temperature(s3) *white blood cells(s3)	0.6755	0.1794	14.1823	0.0002	1.965

(s3) denotes standardized Day 3 variable.
Variable_name1*variable_name2 denotes interaction term.

4.4 Artificial Neural Network model development

A complete description of the architecture of the back-propagation artificial neural network employed in this comparison is presented in the Methods section.

4.4.1 ANN 1: Day 1 artificial neural network

Learning ceased after 123,800 learning epochs. At termination, the minimum average classification error on the developmental data set was 0.0118. The minimum average classification error on the validation set occurred after 23,800 learning epochs and was 0.0875.

4.4.2 ANN 2: Day 3 artificial neural network

Learning on the day three model ceased after 119,200 learning epochs and a minimum average error of 0.0173 was achieved on the 811 patient developmental data set. The minimal average error on the 338 patient validation data set was 0.0701 and was achieved after 19,200 learning epochs.

4.4.3 ANN_{OT} : ANN developed over time

Learning was terminated after 127,600 events with a minimum average developmental error of 0.0013922. The minimal validation set error, achieved after 27,600 learning epochs, was 0.06990.

4.4.4 GenNet: Genetic-algorithm network - day 3 data

Using a Genetic Breeding Pool size of 100, learning continued for 63 minutes before progress was terminated. Termination was set to occur as soon as 20 consecutive generations of learning resulted in a less than one percent improvement in the overall mean square error. The overall mean square error for the developmental training set was 0.023 and the mean square error reported on the validation data set was 0.105.

4.4.5 Consultant's Predicted Outcome

Of the 1,149 patients entered into the study, 308 had an expected duration of stay greater than 72 hours but had either discharge or withdrawal of care scheduled during day three morning rounds. These 308 patients were therefore not eligible for consultant's predictions.

Of the remaining 841 patients, 322 became eligible for consultant's predictions on either Saturday or Sunday. Consultant's predictions could not be obtained prospectively on the weekend and therefore of the total 1,149 patients entered into the study, 519 were

truly eligible for prospective collection of the attending consultant's predicted outcome. Of the 519 truly eligible patients, a consultant's prediction was obtained on 401.

The ICU-based mortality rate for the patients on whom a consultants outcome prediction was available was 21.5 percent and the median length of stay was 7 days (range 3 to 103). Of the patients who qualified for a consultants prediction but on whom a consultant's prediction was not available, the mortality rate was 20.45 percent and the median length of stay was 5 days (range 3 to 179). There were no significant differences between these two groups with respect to these outcomes.

For all 401 patients, the area under the ROC curve for the consultants predictions was 0.8299±0.0285 and the Hosmer-Lemeshow goodness of fit $\hat{C}$ statistic was 36.79 ($p<0.0001$ with 10 degrees of freedom).

4.5 Predictive performance

The primary assessment of predictive performance was conducted by calculating the area under the ROC curve generated by each individual model's application to the same 338 patient validation data set. This validation data set was randomly selected before model development was begun. It was independent of the 811 patient data set used for development and did not directly contribute to the estimation of regression coefficients or network neuronal weights. Table 11 presents the area under the ROC curve and the results of the Hosmer-Lemeshow goodness of fit test for each model applied to the validation database.

Table 11. Area under the ROC curve and goodness of fit of all models assessed on the 338 patient validation database.

	Logistic Regression Models			Artificial Neural Network Models			
	LM 1 (n=338)	**LM 2** (n=338)	**LM Over Time** (n=338)	**ANN 1** (n=338)	**ANN 2** (n=338)	**ANN Over Time** (n=338)	**Genetic Algorithm** (n=338)
AROC ±std err	0.7061 ±0.0395	0.7158 ±0.0395	0.7342 ±0.0385	0.7173 ±0.0391	0.7845 ±0.0362	0.8095 ±0.0347	0.7775 ±0.0366
H-L ($\hat{C}$) p-value	119.8 p<0.001	1350 p<0.001	654.7 p<0.001	29.98 p=0.001	16.0 p=0.098	71.0 p<0.001	46.7 p<0.001

H-L ($\hat{C}$): Hosmer-Lemeshow goodness of fit test $\hat{C}$ statistic chi-square
p-value: probability of H-L gof chi-square with 10 degrees of freedom
aROC: Area under the receiver operating characteristic curve
Std err: standard error

4.5.1 Consultant's predicted outcomes

One hundred and fifty-three of the 401 consultant predictions obtained were on patients who were randomly selected to be in the validation database. The performance of the consultant's predictions was compared to LM_{OT} and ANN_{OT} using this 153 patient validation database (Table 12).

Table 12. Area under the ROC curve and goodness of fit of LM_{OT}, ANN_{OT} and ICU Consultants on a 153 patient validation database.

	Model type		
	Logistic Model Over Time (n=153)	**Neural Network Over Time** (n=153)	**Consultant's Predictions** (n=153)
AROC ±std err	0.6814 ± 0.0518	0.8094 ±0.0442	0.8210 ±0.0432
H-L ($\hat{C}$) p-value	650.17 p<0.0001	7.6815 p=0.659	18.39 p=0.0486

H-L ($\hat{C}$): Hosmer-Lemeshow goodness of fit test $\hat{C}$ statistic chi-square
p-value: probability of H-L chi-square with 10 degrees of freedom
aROC: Area under the receiver operating characteristic curve
Std err: standard error.

4.6 Primary comparisons:

All comparisons were planned *a priori* to address the following primary research questions:

4.6.1. Can artificial neural networks perform mortality prediction in the ICU significantly better than the currently used technique of multivariate logistic regression?

Both LM1 and ANN1 were developed using patient demographic information plus physiologic variables collected on day one of stay. Using the test for correlated areas, there was no significant difference (aROC LM1 0.7061±0.0395 vs. ANN1 0.7173±0.0391, p=0.8953) between the discriminative ability of both approaches.

LM2 and ANN2 were developed using demographic information plus physiologic variables collected on day three of stay. The area under the ROC curves for both models were correlated and ANN2 was found to have significantly better discrimination (aROC LM2 0.7158±0.0395 vs. ANN2 0.7845±0.0362, p=0.0355).

The approach used to develop LM_{OT} and ANN_{OT} was similar in that these models included demographic data, day three physiologic data plus information collected on day one of stay. The area under the ROC curves for these two models was correlated and ANN_{OT} was found to have significantly better discrimination (aROC LM_{OT} 0.7342±0.385 vs. 0.8095±0.0347, p=0.0140).

4.6.2. For patients with a duration of stay over 72 hours, will scoring on day 3 of ICU stay rather than day 1 increase the predictive performance of both artificial neural networks and logistic regression?

LM1 was developed using data available only on day one of stay and LM2 was developed using only day three data. While there appeared to be a marginal increase in area under the ROC curve from the use of day three data, this increase was not statistically significant (aROC LM1 0.7061±0.0395 vs. LM2 0.7158±0.0395, p=0.80).

LM_{OT} used information collected on day one in addition to day three information and appeared to have an improved area under the ROC curve as compared to LM1. This increase in aROC was not statistically significant (aROC LM1 0.7061±0.0395 vs. LM_{OT} 0.7342±0.0385, p=0.5852).

The area under the ROC curve for ANN1 was 0.7173±0.0391. ANN2, which was developed using day three information, had an increased area under the ROC curve at 0.7845±0362, which trended towards statistical significance (p=0.0874). ANN_{OT}, which used all available demographic, day one and day three physiologic data, showed a statistically significant improvement in discrimination over ANN1 (0.7173±0.0391 vs. 0.8095±0.0347, p=0.0598).

4.7 Secondary Comparisons

4.7.1. Can artificial neural networks perform mortality prediction in the ICU significantly better than experienced clinicians, and thus have the potential to become a useful clinical decision support tool ?

Based on a comparison of the area under the ROC curve, the discriminative power of the consultant's predictions was not significantly different from ANN_{OT} (0.8210±0.0432 vs. 0.8094±0.0442, p=0.7684). The consultants were able to discriminate significantly better than LM_{OT} (0.8210±0.0432 vs. 0.6814±0.0518, p=0.0015).

5. Discussion

5.1 Selection of Primary Outcome

When a population of patients under investigation can gain or lose subjects over the course of the study follow-up period, it is called a *dynamic* population. In the special situation where the basic demographic traits of a dynamic population under study remain relatively constant over the period of interest, the dynamic population can be said to be *stable*. In the Intensive Care Unit, the number and basic demographics of the patients requiring treatment remains relatively constant over time even though different patients are continually being admitted and discharged. A population of patients requiring ICU care studied over time can therefore be viewed as *a stable dynamic population*.

A stable dynamic population can be analyzed as a hypothetical fixed cohort if entry into the study corresponds to an event marking the onset of the relevant risk period of interest (*i.e.* study entry defines the beginning of the period during which the subject is a candidate for developing the outcome of interest). [92] In most western countries, intensive care is defined as "concerning itself with the management of patients with life-threatening or potentially life-threatening conditions" and furthermore "such conditions should be compatible with recovery."[93]

According to the principles of outcomes research, to ensure an observational study has *attributional validity,* the patients must be followed at least until this period of increased risk has passed.[94] Attributional validity is concerned with the question of whether or not the care processes under investigation can actually be expected to influence the study outcomes selected.

In Canada, the purpose of the Intensive Care Unit is to support patients through an acute episode of life-threatening illness or injury. It is not the mandate of the ICU to support chronic care patients. Although chronic care patients will be presented to the ICU, it is invariably for an acute complication in their treatment regimen. Since the mandate of the ICU is to resolve acute life-threatening episodes of major illness or injury, it is appropriate to measure the impact of ICU care on an outcome that is directly *attributable* to this acute care process.

Patients are discharged from the ICU when the attending clinician determines that they are no longer at an increased risk of mortality. In this way, a discharge from the ICU can be viewed as a recovery from the original period of risk and thus is inarguably a success of ICU care. A 'failure' in ICU care will result in mortality during a period of time which corresponds with the initial period of high risk. The major assumption behind these definitions is benign and assumes that the attending clinician is highly motivated to keep the patient in the ICU until a true recovery is achieved.

In this study, since entry into the ICU marks the onset of a period of high risk and discharge from the ICU marks the end of that high risk period, we are able to treat our stable dynamic population as a hypothetical fixed cohort and perform analysis on an appropriate outcome that optimizes attributional validity. In other words, recovery/mortality at ICU discharge is a valid outcome even though the follow-up period for each patient is different.

5.2 Patient selection

The patient population admitted to this study was composed of medical, surgical, neurological, cardiac surgery and trauma patients. During the 24 month data collection period, investigators from the CCTC participated in numerous national and multinational randomized controlled trials. In the context of these trials, the severity-of-illness adjusted outcomes of patients admitted to the CCTC were demonstrated to be similar to patients admitted to other tertiary-care teaching ICUs.[95,96]

The 3,728 patients admitted to the ICU during the period of this study accounted for a total ICU stay of 17,080 days. The only restrictive entry criterion included in this project was the requirement that discharge or withdrawal of care had not been scheduled prior to the beginning of morning rounds on calendar day three of stay. While this requirement excluded 68 percent of the total admissions to the CCTC, the 1,181 patients that became eligible for entry into the study accounted for a total ICU stay of 14,088 days. This was 82 percent of the total bed-days used by the ICU during the study period.

Although models were developed using data collected on day one of stay and the predictive performance of these models (LM1 and ANN1) was compared directly to each other, they were not intended to be representative of the broader class of models

developed using day one data (i.e. APACHE III, SAPS II and MPM II). LM1 and ANN1 are highly restrictive and contain only 30 percent of the patients who would be eligible for scoring in a true day one model. The primary purpose of developing and measuring the predictive performance of both LM1 and ANN1 was to determine if these baseline models could be improved by the incorporation of physiologic data that became available over time. The direct comparison of LM1 and ANN1 has internal validity within the constraints of this project, and was intended only to serve as a baseline from which to assess model improvement.

5.3 Data abstraction and variable selection

The reliability of information abstracted from medical records using the APACHE methodology has been shown to have high inter-rater reliability for the APS variables (ICC 0.90) and age (ICC 0.998). The inter-rater reliability for the CHE components has been reported as being much lower (ICC 0.66).[27]

The primary purpose of this project was to compare the predictive performance of two different modeling approaches and not necessarily to compare existing scoring systems nor to develop a totally new scoring system. In order to maximize internal validity, data was collected using a widely accepted methodology. Variables were restricted to those with documented predictive utility and a high degree of inter-rater reliability. The data used for both the pilot project and the current project were abstracted using the APACHE methodology.

The variables collected for the present study included all the physiologic variables indicated for abstraction by APACHE III with two minor modifications: 1) the addition of potassium and; 2) the substitution of hemoglobin for hematocrit. SAPs II identified potassium as a significant predictor of outcome and it is readily available and easy to abstract from the CCTC's patient charts.[55] Although APACHE III includes hematocrit, which is a crude measure of the percent of blood volume composed of red blood cells, hematocrit is not routinely measured in the CCTC. A direct measure of blood hemoglobin content, recorded in grams of hemoglobin per litre of blood, was substituted for the crude measure hematocrit.

The CHE portion of APACHE III has been shown to have the lowest inter-rater reliability. The major demographic components of the CHE were readily available for abstraction from the charts on day three and thus age, gender and source of admission were collected. Both MPM II and SAPS II identify type of surgical admission (elective surgery vs. emergent) as an important predictor of outcome. A comparison of the variables abstracted for this project with those collected by MPM II, SAPS II and APACHE III is presented in Table 13.

Table 13. Comparison of variables collected for this study with MPM II, SAPs II and APACHE III variables.

Variable name	MPM II_{24}	SAPS II	APACHE III	This Study
Temperature		X	X	X
Heart rate		X	X	X
Respiratory rate			X	X
Blood pressure		X	X	X
Hematocrit			X	
Hemoglobin				X
White Blood Cell Count		X	X	X
Albumin			X	X
Bilirubin		X	X	X
Glucose			X	X
Sodium		X	X	X
Potassium		X		X
Bicarb		X		
Creatinine	X		X	X
Blood Urea Nitrogen		X	X	X
Urine output	X	X	X	X
PaO2 and or FiO2	X	X	X	X
PH and PaCO2			X	X
Platelets			X	X
Prothrombin time	X			
Neurologic status	X	X	X	X
Age	X	X	X	X
Source of admission			X	X
Type of admission	X	X	X	X
Vasoactive drug use	X			
Mechanical ventilation	X			
Chronic Health Status	X		X	

5.4 Database validation

Initial screening for irregularities in input variables was undertaken using programs written to filter biologic impossibilities. Upper and lower limits for all physiologic variables were based on an expert consensus process involving three experienced critical care physicians. Each of the clinicians was requested to provide an estimate of an upper and lower limit for all physiologic variables. The physicians were

asked to consider the following question when they established these limits: what is the highest/lowest value of this parameter that you have seen recorded from a patient who was not in eminent risk of death at the time of sampling and was not attributable to a laboratory error or an error in sample processing? The responses from all three clinicians were combined to produce one set of limits for all physiologic parameters. The results of this process were circulated to all three clinicians for final comment. Values in the database that were detected as being out of range, as defined by these consensus- derived limits, were identified and corrected.

In most cases, out of range or missing values in the computer database could be abstracted directly from the study code book. In situations where these values were not available in the code book, the original patient charts were requested from medical records. Of the approximately 300 charts requested, 32 were found to be incomplete, unavailable or missing at time of audit. Although the mortality rate and length of stay observed in these 32 missing patients was not *statistically significantly* different from the remaining 1,149 patients on whom complete information was available, the mortality rate was approximately double that reported in the patients on whom complete information was available.

The purpose of a hospital mortality and morbidity (M&M) rounds is to improve future patient care by learning from past experiences. M&M rounds are usually conducted on patients who undergo catastrophic events, such as mortality. In order to learn from this catastrophic event, a patient's chart is reviewed in detail by a number of different physicians. During review for M&M rounds, the chart would be unavailable from medical records. Since a patient who undergoes an unexpected death in the ICU is more likely to have her/his chart reviewed at M&M rounds and the chart is thus more likely to be unavailable from medical records, this could explain the apparently disproportionate mortality rate observed amongst the patients with missing or unavailable charts.

After completion of the retrospective chart audit, a true missing data rate of 2.62 percent was revealed. This is consistent with both the SAPS II study, which reported missing information in 1.2 percent of all patients[55] and the MPM II developmental data set of 19,442 patients, which contained missing information in 1.6 percent of all cases.[51]

5.4.1 Handling missing values

In the SAPS II and MPM II studies, patients with *any* missing data were excluded from contributing towards the predictive instrument.[51,55] In the APACHE III validation study, imputed averages were used to replace missing values in order to allow these patients to contribute towards the final predictive instrument.[9] Although there are many different ways of imputing missing values, in situations where databases are highly correlated, replacement of missing values may best be achieved by the use of a unique regression equation developed to predict each missing value.[97] Since this entire project is in fact a comparison of two 'regression' techniques (logistic vs. ANN), it was decided that replacement of missing values with imputed averages would provide the most valid baseline comparison of predictive performance.

The two reasons most often cited for neural networks' superior performance when compared with more traditional statistical techniques are: 1) their ability to identify patterns of predictors not detected by standard techniques; and 2) their ability to predict accurately even with noisy or missing input data.[70,98] Although there is some evidence to suggest that neural networks may have the ability to identify novel predictors previously not associated with well investigated outcomes,[71] there is no published evidence to suggest that neural networks can handle missing or noisy information better than logistic regression can.

During the data collection phase of the current project, a second pilot project was undertaken to investigate the performance of ANNs and logistic regression under conditions of excessive missing input data. This second pilot used the predictive models and the 138 patient validation data set developed in the initial pilot project (see Appendix III for complete details of the first pilot).

To create noise in the validation data set, missing values were randomly generated and replaced with imputed average values using a 27 column by 138 row binary transformation matrix. Each element of this matrix was generated using the Bernoulli function in Quattro® Pro for Windows, Version 5.0[a] and had a 95 percent

---

[a] Borland International, Inc., 1800 Green Hills Road, P.O. box 660001, Scotts Valley, CA 95067-0001

probability of assuming a value of one, and a five percent probability of assuming a value of zero. Each element of the validation data set was multiplied by its corresponding element of this transformation matrix. A five percent missing data rate was used because it was double the missing data rate experienced in the first pilot project and thus represented a worst case scenario. Full details of this second pilot project are presented in Appendix VIII.

As reported in the initial pilot project, the ANN and the logistic regression model performed comparably in the 138 patient validation data set (aROC 0.8320 logistic regression vs. 0.8178 for the ANN). In the validation data set with five percent of all elements randomly replaced with imputed average values, both the logistic regression and the ANN reported marginally reduced performance (aROC 0.804 for logistic regression vs. 0.800 for the ANN). Since the performance of both the logistic regression and the ANN models responded similarly to the introduction of noise into the inputs, it was decided that handling missing values in the definitive project by substitution of imputed average values would not likely account for any observed differences in performance between the two different modeling approaches.

5.5 Logistic regression model development

5.5.1 Basic logistic regression modeling methodology

In order to make the logistic regression modeling methodology transparent and repeatable, each step and decision point was outlined in advance.[84] The explicit steps used to develop the logistic regression model are outlined below:

Step 1. Variable Selection

Step 2. Specification of the Maximum Model

Step 3. Assessment for Multicollinearity

Step 4. Model Selection

Step 5. Regression diagnostics

Variable selection, or model entry criteria, was set to the more liberal threshold of $p<0.25$ for entry into the maximum model to improve the possibility of including important confounding variables in the model development process.[82]

In this project, categorical demographic variables were represented as zero cell referent dummy variables. Demographic and physiologic variables were treated as separate chunks during the specification of interaction terms eligible for the maximum model. While all physiologic variables were considered in interaction terms, interactions between demographic and physiologic variables were not considered. The assumption of a lack of significant interactions between demographic and physiologic variables is based on the finding that the APACHE III, the MPM II and the SAPS II models did not detect significant demographic-physiologic interactions.[9,51,55]

As a general principle, reliability in the estimation of the coefficients in any regression model decreases as the model becomes more complex.[92] Based on the results of the pilot project, problems with multicollinearity were anticipated and parsimonious model building strategies that would reduce the potential for multicollinearity were embraced early in the model development process. Under these principles of parsimony, three-way interaction terms were not considered for entry into the maximum model.

One of the major limitations of the backward elimination procedure is that it may be more prone to problems associated with singularities in the information matrix than is the forward selection process.[92] To avoid this problem, collinearity was assessed using Eigenanalysis before the modeling process was begun and as a result, all continuous variables were standardized. Standardization (calculation of a z-score) was selected over simple centering since it addresses issues associated with multicollinearity *and* scaling problems. Scaling of the artificial neural network inputs was required in the pilot project in order to achieve convergence (Appendix III).

Standardization is a simple linear transformation of the input variables and as such, does not alter the estimated coefficients or estimated information matrix obtained from a maximum likelihood procedure. Any apparent reduction in multicollinearity achieved by centering or scaling is actually only an artifact when using an estimation procedure that is inherently independent of linear transformations of the independent variables.[99]

Numerous goodness of fit statistics that were originally developed to assess least-squares linear regression models have been extended to assess logistic regression models.[100,101] The two most widely used goodness of fit statistics that were developed

specifically for logistic regression are the Hosmer and Lemeshow $\hat{H}$ statistic and the Hosmer and Lemeshow $\hat{C}$ statistic.[85,82] Both of these Hosmer-Lemeshow tests are based on grouping the estimated probabilities into deciles of risk and then calculating a summary statistic that follows the chi-square distribution.

The major difference between the $\hat{C}$ and the $\hat{H}$ statistics lies in the approach used to create the groupings. Although both tests have desirable properties and seldom yield conflicting results, because it does not balance group sizes, the $\hat{H}$ statistic becomes unstable if the number of cases in any one particular grouping approaches zero.[48,100,101] For this reason, the Hosmer-Lemeshow $\hat{C}$ statistic was selected to assess the goodness of fit of both the logistic regression and ANN models throughout this project.

5.6 Artificial neural network model development

Back-propagation neural networks are still the most common type of neural network evaluated by medical researchers,[75,102,103] but some researchers have begun to investigate the application of other network architectures. One such investigation compared the performance of a neural network developed using a genetic learning algorithm to the performance of logistic regression.[79] Both the genetic learning network and the logistic regression model were developed to predict outcome in intensive care unit patients and the investigators reported the genetic learning ANN outperformed the logistic regression model based on area under the ROC curve in a validation data set (0.863 vs. 0.753). Because this finding demonstrated the potential utility of genetic ANNs, we undertook a direct comparison of a genetic learning ANN to a back propagation ANN. Other researchers who have compared back-propagation networks to alternative architectures have demonstrated the back-propagation architecture to be superior.[104]

The genetic learning ANN and the back-propagation network were developed using information collected on day three of stay and their predictive performance was compared on the 338 patient validation data set. Although the back-propagation network did not perform significantly better than the genetic learning network based on a direct comparison of area under the ROC curves (0.7845 vs. 0.7775, p=0.4297), the back-propagation network demonstrated good fit on the validation data set (H-L gof, p=0.098)

whereas the genetic learning net did not (H-L gof, $p<0.001$). Based on the finding that the genetic learning network failed to show significant improvements over the back-propagation network, the back-propagation network was selected as the topology to be evaluated throughout the remainder of the project.

The back-propagation networks developed for evaluation in this project were constructed using widely accepted approaches and algorithms that are published in detail elsewhere.[105]

5.7 Feasibility

5.7.1 Time frame for completion

The pilot project was conducted using a database collected from August 5, 1991 to February 5, 1992 in the Richard Ivey CCTC. During that 6 month period, 614 patients stayed at least 24 hours and 422 were admitted for at least 72 hours. The average length of stay was 5.67 days with a standard deviation of 7.68 days. Based on this pilot information, it was estimated that the minimum required sample size of 1,200 patients would be achieved after 18 to 24 months of data collection.

The minimum estimated sample size of 1,200 patients was achieved after 24 months of data collection. During the period of the study, the Ministry of Health reduced the funding of the Richard Ivey CCTC from a 30 bed to a 26 bed ICU. This reduction in the number of funded beds resulted in a marginally shorter average length of stay (4.58 ± 10.4 days) compared to that of the pilot project. The original study design called for data collection to continue for 24 months to account for optimism in the original recruitment rate estimates. Thus the minimum estimated sample size was achieved in the *a priori* planned data collection period.

5.8 Primary comparisons:

The artificial neural network models developed using day three data (ANN2 aROC=0.7845) and the data collected over time (ANN_{OT} aROC=0.8095) performed marginally significantly better than the logistic regression models developed using the same available data (LM2 aROC=0.7158 and LM_{OT} aROC=0.7342). The only neural network model that did not outperform its logistic regression counterpart was the day one net (LM1 aROC=0.7061 vs. ANN1 aROC=0.7173).

On initial inspection it might appear that both the day one logistic regression and ANN models displayed relatively poor predictive performance on the validation database as compared to other models developed using day one information. On validation data sets, the MPM II model performed with an area under the ROC curve of 0.824 and the SAPS II model performed with an area under the ROC curve of 0.86.[51,55] It is important to note however, that both ANN1 and LM1 were used to predict outcomes on patients who had a minimum duration of stay of at least 72 hours.

In the most rigorous study to demonstrate the influence of length of stay on the power of discrimination, the performance of a day one SAPS logistic regression model was shown to decrease in direct relationship to length of stay (Figure 1). The area under the ROC curve for this SAPS model showed a statistically significant decrease from 0.79 on day one to 0.72 in patients whose minimum duration of stay was three days.[59] Given that SAPS, LM1 and ANN1 are being evaluated in long stay patients(>3 days), the area under the ROC curves for all three models is remarkably similar (0.72 SAPS, 0.70 LM1 and 0.72 ANN1).

Both the MPM II and the APACHE III investigators attempted to improve the performance of their respective predictive models in long stay patients. The MPM II models that were developed to adjust for changes over time demonstrated a progressive decrease in the area under the ROC curve from MPM II_{24} to MPM_{72} (MPM II_{24} aROC=0.836, MPM II_{48} aROC=0.796 and MPM II_{72} aROC=0.752).[52] The APACHE III time dependent models also demonstrated a progressive decrease in area under the ROC curve from 0.90 for the day one model, to 0.88 for the day 3 model and 0.84 for the day 15 model (Table 1).[58]

The loss in discrimination in the MPM II logistic regression models was thought to reflect a true loss in discriminative potential inherent in the data. The authors suggested that as patients remained in the ICU for longer periods of time, MPM II demonstrated a progressive loss of discriminative power because either a) the patients' conditions simply became too complex for accurate predictions to be obtained or b) since they failed to respond to treatment, their physiological parameters became uninformative. It was also demonstrated that after ICU admission, large numbers of patients either

recovered enough to be transferred out of the ICU relatively quickly or they suffered massive injury and died in a relatively short time.[52]

In this current project, although the performance of the logistic regression models did not improve significantly as more information was included over time, (LM1 aROC=0.7061 to LM_{OT} aROC=0.7342), they did not demonstrate a progressive loss of discriminative power. The artificial neural network models actually demonstrated marginally significantly improved discrimination (ANN1 aROC=0.7173 to ANN_{OT} aROC=0.8095, p=0.0598) when data available on day three of stay was used to update the model.

Within the constraints of this project, back-propagation ANNs were able to predict patient outcomes better than logistic regression models. When these artificial neural network models were updated as more information became available over time, they were able to significantly improve their power to discriminate between patients who would subsequently live or die. There are four possible reasons that may explain these findings:

1. ANN methodology can identify predictors that statistical techniques do not;
2. ANNs implicitly detect all possible interaction terms;
3. ANNs can identify complex nonlinear relationships; and
4. ANNs are insensitive to problems associated with multicollinearity.

5.8.1 Improved ability to identify predictors

In previous research, neural networks have been shown to place importance on clinical signs not previously thought to be of diagnostic importance. For example, an assessment of a neural network that was developed to aid in the diagnosis of acute myocardial infarction placed importance on the presence of rales on auscultation to improve its diagnostic accuracy.[71] The presence of rales was not previously recognized in the cardiology literature as a diagnostic sign of myocardial infarction.

The APACHE III, SAPS II and MPM II models all include the basic components of the GCS score as an important predictor of outcome. The GCS score is a composite index of global neurological status composed of assessments of eye opening, motor response and verbal ability.[106]

In this study, univariate analysis of the GCS collected on day one of ICU stay found that it was not a significant predictor of outcome in patients with a minimum length of stay of three days (p=0.9263). Even using the liberal inclusion criteria of a univariate probability value less than 0.25 set for this project, GCS could not be included in the maximum model for LM1.

According to neural network theory, all possible predictors are presented as inputs to the model under development. The back-propagation algorithm uses progressive feedback and error correction cycles to adjust numerical weights between the input variables in order to maximize the overall predictive accuracy of the network. If an input variable has low predictive value, the pathways that connect it to other neurons will be assigned low numerical weights and thus it will play a small role in determining outcome. One way to assess the relative contributions of each input variable is to sum all connection weights in the network and compare the amounts directly attributable to each input variable.[71]

Based on this contribution factor analysis, the most important predictor used by the day one neural network was blood urea nitrogen, with a contribution factor of 0.05258, followed closely by albumin at 0.04564 and the Glasgow Coma Scale score at 0.04562. Thus, the neural network developed using the day one data identified GCS as the third most important predictive variable whereas classical statistical methodology did not even identify GCS for inclusion in the maximum model.

The GCS was originally developed and validated in a population of head trauma patients for scoring on initial presentation.[106] Independent research has shown that GCS scored on the day of admission accurately predicts outcomes in head trauma patients but as length of stay increases, the day one GCS loses its predictive power.[43] Since patients admitted to the CCTC were a mixed population of general ICU patients, and not specifically head trauma patients, it is not surprising that admission GCS was found to be a poor predictor in this study: it was being used in a population of patients different from its development population and even in head trauma patients, admission GCS is known to have reduced predictive power in long stay patients.

The importance placed on GCS by the ANN would tend to argue that admission GCS has some predictive value in long stay patients when combined with other

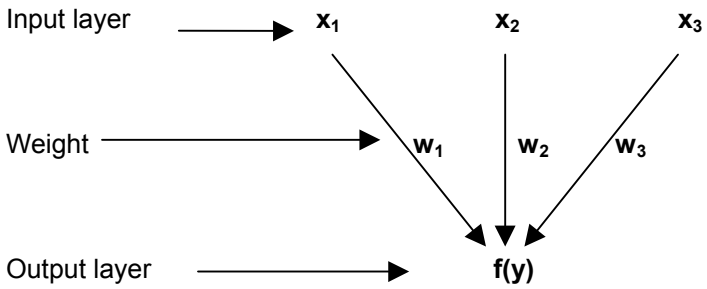
physiological terms in the context of an interaction. Although the day one ANN did not outperform the day one LM, the weight placed on GCS by ANN1 serves as an excellent example of how neural network methodology can identify predictors that classical statistical approaches may not.

The use of a univariate p-value of 0.25 to determine eligibility for entry into the maximum logistic model is a somewhat arbitrary design constraint placed on the regression methodology by the authors. It is possible that a model built using a forward stepwise approach or one that used alternative criteria for eligibility would have detected the predictive importance of GCS. Previous ICU-based predictive models have used univariate screening with a p-value of 0.10 for inclusion.[51,55] Employing an approach that used univariate screening with a p-value of 0.25 for inclusion in the logistic regression model was considered representative of accepted practice.

5.8.2 ANNs automatically detect all possible interactions

In the simplest form of ANN, all the variables in the input layer map directly to the outcome layer using a simple linear mapping function (Figure 3). The coefficient, or weight, associated with each individual input variable is determined using a pre-specified algorithm optimized to minimize predictive error. One possible approach to determining this weight is to use a modified least-squares algorithm. These weights or coefficients are then combined and mapped onto the final outcome using a non-linear transfer function.

Figure 3. Diagrammatic representation of a simple neural network



Where x_1,x_2,x_3 are the input variables,
w_1,w_2,w_3 are the modifiable weights and
f(y) is the logistic output function.

This simple network can be represented mathematically by Equation 2 and can be seen to be algebraically equivalent to a regression equation containing main effects only.

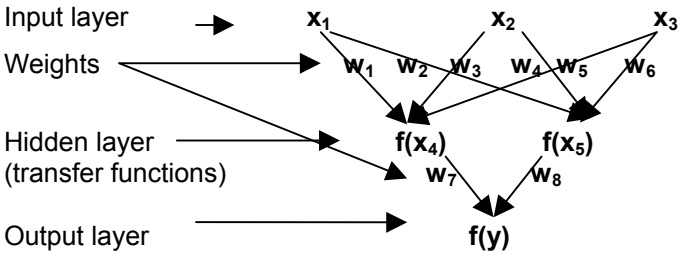
Equation 2. Algebraic representation of a simple neural network.

$$y = w_1x_1+w_2x_2+w_3x_3$$

Where x_1,x_2,x_3 are the input variables,
y is the output and
w_1,w_2,w_3 are the modifiable weights.

If a single hidden layer is inserted between the input layer and the output layer of this simple network, the number of interconnections increases dramatically (Figure 4).

Figure 4. Diagrammatic representation of a neural network with one hidden layer



where x_1,x_2,x_3 are the input variables,
f(y) is the logistic output function,
$w_{1\text{-}8}$ are the modifiable weights and
$f(x_4)$ and $f(x_5)$ are the hidden layer transfer functions.

The algebraic representation of a two layer network is much more complex, and depends on the specific architecture of each network but in general terms, can be represented by Equation 3.

Equation 3. Algebraic representation of single layer neural network

$$y = w_7f(x_4)+w_8f(x_5),$$

$$x_4=w_1x_1+w_3x_2+w_4x_3,$$

$$x_5=w_2x_1+w_5x_2+w_6x_3.$$

where x_1,x_2,x_3 are the input variables,
y is the output,
$w_{1\text{-}8}$ are the modifiable weights and
$f(x_4)$ and $f(x_5)$ are the hidden layer transfer functions.

As a function of the complexity of the interconnections between each neuron, and the use of nonlinear transfer functions within the hidden neurons, an ANN with one hidden layer *automatically* models two-way and higher-order interaction terms.[62] It must be recognized that logistic regression *can* model higher order interaction terms and a diligent investigator *could* detect all possible interactions using a logistic regression-based modeling methodology but when logistic regression is used, interactions of interest must be pre-specified.

The fact that the specific methodology chosen for this study does not include all possible interactions for inclusion in the LM was a design decision and may have limited the generalizability of the findings. Previous ICU-based investigations have consistently failed to demonstrate the importance of three-way or other complex interaction terms.[9,51,55] Given considerations of parsimony, a mildly restrictive approach to the investigation of higher-order interaction terms was deemed reasonable.

Because of the ease with which ANNs model interaction terms, some authors have suggested that different network architectures could be applied to data sets specifically to detect the presence of higher-order interaction terms.[62] Once detected, these interactions could be investigated using classical techniques and included in any statistical model building process.

5.8.3 ANNs automatically consider complex nonlinear relationships

In order to maintain homeostasis, physiological systems must exhibit complex adaptive control systems behavior. For example, the delivery of oxygen throughout the body is regulated by sensors and feedback mechanisms that constantly monitor demand and adjust stroke volume, heart rate, minute ventilation, regional blood flow, blood pressure, organ blood flow and even flow within and between capillary beds. The exact means used to adjust the system under any given demand is decided by a delicate interplay of an incredible number of feedback control mechanisms. In a critical illness such as sepsis, the interplay between these control mechanisms can become perturbed in such a way that homeostatic control is lost and extremely complex patterns of responses, such as shock or pathologic supply dependency, are demonstrated.[107]

People who are critically ill and require intensive care are by definition, exhibiting a pathologic physiological response to some external or internal stimulus. Since homeostasis is often lost and because of the complexity of the control systems in the human body, one should not expect these pathologic responses to be linear and predictable. The database collected for this project is composed of physiologic variables that could be expected to demonstrate complex, perhaps nonlinear interrelationships between each other and with eventual outcome.

In Figure 4 and Equation 3, the terms $f(x_4)$ and $f(x_5)$ refer to representative hidden layer transfer functions and $f(y)$ refers to the output function of a representative two layer

ANN. In a network that models a binary outcome, f(y) is typically set to the logistic function but $f(x_4)$ and $f(x_5)$ can be set to any number of nonlinear transfer functions (Figure 2).[74] Previous authors have demonstrated that a multi-layered back-propagation network with nonlinear transfer functions can be used to model the behavior of different nonlinear systems.[108] Although a three-layered network is required to approximate a system with polyhedral defining equations,[109] in general a two-layered network is adequate to approximate most nonlinear systems.[105] Since a two-layered network was evaluated in the pilot project and since it demonstrated acceptable performance, two-layered networks were evaluated in this current project.

The ability of a neural network to detect nonlinear relationships is a property of the type of nonlinear transfer function assigned to each neuron layer. By adjusting the weights assigned to the interconnections between each neuron, the back-propagation algorithm can optimize the use of any specific transfer function. Thus if a significant degree of nonlinearity exists between the input and the outcome variables, the back-propagation algorithm will automatically adjust the weights between neuron layers to reflect this relationship and because detection of these relationships is automatic, the end user does not have to search for or specify higher order polynomials.[75]

One possible deficiency in the logistic regression modeling methodology evaluated in this study is the lack of nonlinear terms as main effects. It is possible that the overall explanatory power of the model could have been improved by the fitting of a square, log or polynomial term. Alternately, a cubic splines regression could have been used to determine cut-points for continuous variables such that a zero-cell referent dummy variable could be used to model a nonlinear relationship.[97] While there is extensive evidence documenting the nonlinear relationships between physiological parameters measured in the ICU,[110] all the published logistic regression models based on physiological parameters demonstrate that acceptable performance can be achieved in models without nonlinear terms.[9,51,55]

Since this current project did not formally investigate the impact of nonlinear terms in the regression model, a follow-up project should be designed. The reader is cautioned that the results of this current project should not be generalized to logistic

model building approaches considering nonlinear terms until subsequent research has been undertaken.

5.8.4 ANNs are insensitive to problems associated with multicollinearity

In the initial attempt at developing a logistic regression model using the data that became available over time, the model LM_{OT} was first developed using *all* physiological terms that were shown to be important univariate predictors on day one and on day three (alb, bun, creat, dbp, FiO2, glu, hr, PaO2, pH, pts, rr, uo, wbc,gcs3, bun3, creat3, dbp3, FiO23, hr3, na3, PaCO23, PaO23, pH3, pts3, rr3, temp3, uo3 and wbc3 *plus* all two-way interactions). The presence of severe multicollinearity was detected in the maximum model using Eigenanalysis (condition index 584.74) and correction was attempted through standardization. After standardization, Eigenanalysis did not detect the presence of multicollinearity (condition index 14.62) and the maximum model was presented to the backwards selection process. When the maximum model was presented to SAS, the estimates of the regression coefficients in the SAS output exhibited instability consistent with the presence of multicollinearity and the backwards elimination process was aborted.

This true multicollinearity was most likely a result of the complex time-dependent relationships within the data. To avoid the negative effects of true multicollinearity due to time-dependency, numerous design options are available. For example, problems associated with multicollinearity could be reduced by using a forward selection approach with forced entry of a limited number of main effects. Previous researchers have reduced the impact of true multicollinearity by compressing all or part of the available input variables using the expected probability output from a previous model as an input to the current model.[58] The final LM_{OT} presented in this project contained the predicted output of LM1 as an input variable instead of the individual day one variables and did not exhibit any problems consistent with the presence of true multicollinearity.

ANN_{OT} was presented with the same variables originally investigated for LM_{OT} (alb, bun, creat, dbp, FiO2, glu, hr, PaO2, pH, pts, rr, uo, wbc,gcs3, bun3, creat3, dbp3, FiO23, hr3, na3, PaCO23, PaO23, pH3, pts3, rr3, temp3, uo3 and wbc3). Experience with the logistic modeling process demonstrated the presence of multicollinearity that could not be corrected using scaling or standardization techniques. In order to obtain valid

estimates using logistic regression, a design approach that uses data compression was required. The neural network was insensitive to the presence of the true multicollinearity in this data set and did not exhibit any problems converging on a solution.

5.9 Relevance of findings

5.9.1 ICU management applications

Accurate day one logistic regression models allow managers to compare outcomes between different ICUs through the calculation of observed to expected mortality ratios.[111] This validated approach to outcome comparison is a necessary requirement to allow the process and structure investigations that serve as a basis for quality assurance projects and resource allocation decisions.[112]

Previous research has demonstrated that the accuracy of logistic regression-based admission scores decrease as a patient's length of stay in the ICU increases.[52,57,59] Although attempts have been made to improve the predictive performance of logistic regression-based models in long stay patients, the resultant time dependent models have not been able to improve performance significantly over the initial day one models.[58] In this project however, the time-dependent ANN developed on long stay patients performed significantly better than a time-dependent logistic regression model.

The finding that time-dependent ANNs can lead to improved predictive performance in long stay patients is not just of academic interest. While the patients included in this project accounted for only 32 percent of all admissions to the ICU, they accounted for 80 percent of the total ICU bed use (17,080 total bed days, 14,088 attributable to patients in this study). Since length of ICU stay is the single most important surrogate measure of costs,[113] the cohort of patients targeted for entry into this study accounted for approximately 80 percent of the total ICU costs.

The inventory management rule of '80-20' states that 20 percent of any business' inventory usually accounts for 80 percent of its holding costs. Good inventory managers use this knowledge to attempt to identify this high-cost 20 percent and target it for special management attention. The improved outcome prediction of time-dependent ANNs could allow ICU managers to increase the accuracy of severity adjusted outcome comparisons in an economically important sub-group of patients. Hopefully this increased accuracy

will lead to a better understanding of the process and structure measures that lead to improved patient care from the perspective of the unit manager, the care deliverer, and the patient.

5.9.2 Clinical utility

In this project, the predictive power of the time-dependent ANN was compared directly to that of senior ICU consultants. Using the 153 patients in the randomly selected validation database on whom consultant predictions were available, no significant difference between the area under the ROC curves was found (ANN_{OT} aROC 0.8094 vs. 0.8210). However, the consultant's predictions demonstrated a significantly poor fit to the data (H-L gof p=0.0486); whereas the ANN_{OT} demonstrated good fit (H-L gof p=0.659).

The ANN is an objective method of evaluating patient outcomes. It is probably unreasonable to expect the ANN alone could support individual patient level treatment decisions. Since the ANN did demonstrate performance similar to *experienced* clinicians, it is possible that ANNs could prove useful to clinicians-in-training. In the ICU, clinicians-in-training communicate daily with relatives of patients and are often asked for estimates of 'chances of survival'. The ANN could prove a useful tool in supplementing or guiding their clinical judgement in communicating these estimates to patient's families.

Improvements in network performance may be achieved by including the clinician's risk estimates as an input variable. Perhaps by presenting the networks output to the attending clinician as a 'second opinion', the clinician may feel more comfortable making certain decisions. Comparing the performance of ANNs to that of experienced clinicians was a secondary objective of this project and as such, these findings definitely warrant further investigation before they are considered for applications in any clinical situation.

5.10 Achieving improved performance with ANNs

This project discussed four inherent properties of neural networks: 1) ANN methodology can identify predictors that statistical screening does not; 2) ANNs

implicitly detect all possible interaction terms; 3) ANNs can identify complex nonlinear relationships; and 4) ANNs are insensitive to problems associated with multicollinearity.

In order to give an ANN the opportunity to identify novel predictors, care should be taken to present it with input variables that include *all possible predictors*. Projects that present ANNs with inputs screened based on statistical principles may not realize the full potential of using an ANN.

The type of problems where networks hold the most promise of acceptable performance tend to be complex, have nonlinear interrelationships with outcome, and have the potential for many different types of interaction terms. Although statistical techniques can model complex nonlinear relationships, great care must be taken to identify and include all appropriate terms. When investigating high order polynomials and interaction terms, the statistical researcher is often forced to apply the principles of parsimony in order to avoid the introduction of problems associated with multicollinearity.

It is important to note that the time-dependent network developed in this project was presented with an extremely complex data set which included at least 13 pairs of physiological variables collected over time. These variables were standardized and entered directly into the ANN and despite a high degree of interrelationships, the ANN did not demonstrate any problems associated with multicollinearity. It is clear that ANNs are stable under situations where there is a high degree of correlation between the input variables.

It is likely that ANNs will demonstrate acceptable performance when they are applied to problems that allow them to capitalize on all four of these inherent properties. ANNs should be investigated when the problem is complex, nonlinear and contains the possibility of severe multicollinearity in the inputs. When applied to these types of problems, ANNs should be presented with all possible predictors as inputs to fully capitalize on their potential to deliver optimal predictive performance.

6. Summary

In this project, data were collected on 1,149 consecutive ICU admissions who were present at morning rounds on calendar day three of ICU stay. A series of representative logistic regression models was developed using a backwards elimination model building process.[82,84] The performance of these logistic regression models was compared to a corresponding series of back-propagation ANNs.

Initial baseline models were developed using data collected on day one of stay (LM1 and ANN1). A second pair of models (LM2 and ANN2) was developed using data collected on day three of stay and a third pair was developed to incorporate changes over time (LM_{OT} and ANN_{OT}) using a combination of day one and day three data.

Although all three LMs demonstrated good fit and calibration on the 811 patient developmental database, they demonstrated poor fit (H-L $p<0.001$) on the 338 patient validation data set. The predictive performance of the LM based on day three data was not significantly different from the predictive performance of the baseline day one LM (aROC LM1=0.7061 vs. LM2=0.7158, p=0.80). Incorporating changes over time into the LM by using day one and day three information also did not improve predictive performance (aROC LM1=0.7061 vs. LM_{OT}=0.7342, p=0.5852).

Compared to the ANN developed using admission data, ANN2 demonstrated a trend towards a significant improvement in predictive performance (aROC ANN1=0.7173 vs. ANN2=0.7845, p=0.0874). By including both day three and day one data, ANN3 actually demonstrated significantly better performance than ANN1(aROC ANN1=0.7173 vs. ANNOT =0.8095, p=0.0598). Both ANN2 and ANN3 also performed significantly better than their corresponding LM models (aROC LM2=0.7158 vs. ANN2=0.7845, p=0.0355 and aROC LM_{OT}=0.7342 vs. ANN_{OT}=0.8095, p=0.0140).

There are four inherent properties of ANNs that could explain their ability to predict better than LM in certain situations. First, it is well established in the literature that ANNs can identify and place importance on predictors that classical statistical techniques do not. Second, ANNs were designed to automatically identify and include complex nonlinear relationships. Third, ANNs were also designed to implicitly detect

and emphasize all possible interaction terms and finally, we demonstrated that ANNs are inherently insensitive to problems associated with multicollinearity.

The ability to accurately predict outcomes in patients with an ICU stay greater than three days has been a major limitation of logistic regression-based admission models. The finding that ANNs can lead to improved predictive accuracy in this economically important sub-group of patients could lead to improved insight into the process and structure measures associated with optimal care.

Although these findings are novel and potentially extremely important, it must be noted that this study optimized internal validity at the expense of potential generalizability. It is suggested that a number of confirmatory studies need to be undertaken before these findings can be generalized and applied elsewhere.

7. Future directions for research

7.1 ICU management applications

The finding that ANNs delivered improved predictive performance in patients with a duration of ICU stay greater than three days should be re-examined using a more representative database. This database should contain demographic and daily physiologic information on consecutive admissions from a representative sample of intensive care units.

The previous study compared the performance of ANNs to a LM developed over time using a widely accepted approach to logistic regression model development. It is suggested that in the next study, the performance of an ANN developed using the methodology outlined in this thesis be compared to the predictive performance of a validated ICU scoring system, such as the APACHE III or MPM II score. More importantly, the ANN developed using data collected over time should be compared with the performance of the daily APACHE III predictive system or the MPM II_{24} or MPM_{72}. By performing this comparison, the potential improvement over an accepted risk stratification system could be evaluated.

There is no need for this study to be prospective in nature as there are at least 3 databases (MPM II, SAPS II and APACHE III) in existence that could satisfy the requirements for such a study.

7.2 Clinical utility

If the performance characteristics of the new ANN model developed on a more representative database appears to offer a clinically significant improvement over currently available scoring systems, a series of prospective clinical studies could be undertaken to investigate potential uses and roles. After an appropriate model has been developed and assessed using a database composed of a representative sampling of ICUs, a series of prospective studies should be undertaken to compare its predictive performance to that of expert clinicians and clinicians in training. The use of the clinician's predictions as an input variable could be investigated as a means of further improving predictive performance.

Previous research has demonstrated that the use of a validated predictive tool to support clinical decisions has been somewhat controversial. It is possible that using a combination of clinical judgement, artificial neural networks and/or logistic regression to investigate the clinical decision process, insights could be gained that could translate into improvements in the efficiency of care. It is this author's recommendations that management applications of ANNs be thoroughly understood and validated before any further research into clinical applications be undertaken.

8. References

[1] Knaus W, Wagner D and Draper E. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Development of APACHE. *Crit Care Med* 1989;**17**(12 Pt 2):S181-185.

[2] Schneiderman LJ, Jecker NS and Jonsen AR. Medical futility: its meaning and ethical implications. *Ann Intern Med* 1990;**112**:948-954.

[3] Griner PF. Medical intensive care in the teaching hospital: Costs versus benefits: the need for an assessment. *Ann Intern Med* 1972;**78**:581.

[4] Civetta JM: Selection of patients for intensive care. *In: Recent Advances in Intensive Therapy*, Number 1. Ledingham, ImcA (Ed). Edinburgh: Churchill Livingstone, 1977.

[5] McPeek B, Gilbert JP and Mosteller F. The clinician's responsibility for helping to improve the treatment of tomorrow's patients. *N Engl J Med* 1980;302:630.

[6] Knaus WA, Draper EA and Wagner DP. Toward quality review in intensive care: the APACHE system. *Qual Rev Bull* 1983;**9**(7):196-204.

[7] Seneff M and Knaus WA. Predicting patient outcome from intensive care: a guide to APACHE, MPM, SAPS, PRISM and other prognostic scoring systems. *J Intensive Care Med* 1990;**5**:33-552

[8] Rutledge R, Fakhry SM, Rutherford EJ, Muakkassa F, Baker CC, Koruda M, *et al*. Acute Physiology and Chronic Health Evaluation (APACHE II) score and outcome in the surgical intensive care unit: an analysis of multiple intervention and outcome variables in 1,238 patients. *Crit Care Med* 1991;**18**:1048-1053.

[9] Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, *et al.* The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991;**100**(6):1619-36.

[10] Muller JM, Herzig S, Halber M, Stelzner M and Thul P.[The acute physiology score as a stratification and prognostic criterion in patients in a surgical intensive care ward]. *Chirurg* 1987;**58**(5):334-40.

[11] Ferraris VA and Propp ME. Outcome in critical care patients: A multivariate study. *Crit Care Med* 1992;**20**:967-976.

[12] Chang RWS, Jacobs S and Lee B. Predicting outcome among intensive care unit patients using computerized trend analysis of daily Apache II scores corrected for organ system failure. *Intensive Care Med* 1988;**14**:558-566.

[13] Civetta JM, Hudson-Civetta JA, Kirton O, Aragon C and Salas C. Further appraisal of APACHE II limitations and potential. *Surg Gynecol Obstet* 1992;**175**:195-203.

[14] Hinton GE. How neural networks learn from experience. *Sci Am* 1992;267:145-151.

[15] Waibel A and Hampshire J. Building blocks for speech: Modular neural networks are a new approach to high-performance speech recognition. *Byte* 1989;**14**:235.

[16] Nightingale C, Myers DJ and Linggard R. *Neural Networks for Vision, Speech and Natural Language.* London: Chapman and Hall, 1992.

[17] Maclin PS and Dempsey J. Using an artificial neural network to diagnose hepatic masses. *J Med Syst* 1992;**16**:215-225.

[18] Scott JA and Palmer EL. Neural network analysis of ventilation-perfusion lung scans. *Radiology* 1993;186:661-664.

[19] Goldberg V, Manduca A, Ewert DL, Gisvold JJ and Greenleaf JF. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. *Med Phys* 1992;**19**:1475-1481.

[20] Lemeshow S, Teres D, Avrunin JS and Pastides H. A comparison of methods to predict mortality of intensive care unit patients. *Crit Care Med* 1987;**15**(8):715-722

[21] Lemeshow S and Le Gall JR. Modeling the severity of illness of ICU patients. A systems update. *JAMA* 1994 Oct 5;272(13):1049-1055

[22] Knaus WA, Zimmerman JE, Wagner DP, Draper EA and Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med* 1981;**9**(8):591-7.

[23] Wagner DP, Knaus WA and Draper EA. Statistical validation of a severity of illness measure. *Am J Public Health* 1983;**73**(8):878-84.

[24] Wagner DP, Draper EA, Abizanda-Campos R, Nikki P, Le-Gall JR, Loirat P *et al*. Initial international use of APACHE. An acute severity of disease measure. *Med Decis Making* 1984;4(3):297-313.

[25] Wong DT and Knaus WA. Predicting outcome in critical care: the current status of the APACHE prognostic scoring system. *Can J Anaesth* 1991;**38**(3):374-83.

[26] Knaus WA, Draper EA, Wagner DP and Zimmerman. APACHE II: A severity of disease classification system. *Crit Care Med* 1985;**13**:818-830.

[27] Damiano AM, Bergner M, Draper EA, Knaus WA and Wagner DP. Reliability of a measure of severity of illness: acute physiology of chronic health evaluation--II. *J Clin Epidemiol* 1992;**45**(2):93-101

[28] Knaus WA, Draper EA, Wagner DP and Zimmerman JE. An evaluation of outcome from intensive care in major medical centers. *Ann Intern Med* 1986;**104**(3):410-8.

[29] Oh TE, Hutchinson R, Short S, Buckley T, Lin E and Leung D. Verification of the Acute Physiology and Chronic Health Evaluation scoring system in a Hong Kong intensive care unit. *Crit Care Med* 1993;**21**(5):698-705.

[30] Berger MM, Marazzi A, Freeman J and Chiolero R. Evaluation of the consistency of Acute Physiology and Chronic Health Evaluation (APACHE II) scoring in a surgical intensive care unit. *Crit Care Med* 1992; **20**(12):1681-7.

[31] Zimmerman JE, Knaus WA, Judson JA, Havill JH, Trubuhovich RV, Draper EA *et al*. Patient selection for intensive care: A comparison of New Zealand and United States hospitals. *Crit Care Med* 1988;**16**:318-326.

[32] Sirio CA, Tajimi K, Tase C, Knaus WA, Wagner DP, Hirasawa H *et al*. An initial comparison of intensive care in Japan and the United States. *Crit Care Med* 1992;**20**(9):1207-15.

[33] Nouira S, Belghith M, Elatrous S, Jaafoura M, Ellouzi M, Boujdaria R, *et al*. Predictive value of severity scoring systems: comparison of four models in Tunisian adult intensive care units. *Crit Care Med* 1998;**26**(5):852-9

[34] Wong DT, Crofts SL, Gomez M, McGuire GP and Byrick RJ. Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients. *Crit Care Med* 1995;**23**:1177-1183.

[35] Sackett DL, Haynes RB, Guyatt GH and Tugwell P. *Clinical Epidemiology: A basic science for clinical medicine. 2nd Ed.* Boston, MI: Little, Brown and Company, 1991.

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

[37] Knaus W and Wagner D. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Individual patient decisions. *Crit Care Med* 1989;**12**(2):S204-9.

[38] Knaus W, Draper E and Wagner D. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Introduction. *Crit Care Med* 1989;**12**(2):S176-80.

[39] Wagner D, Draper E and Knaus W. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Analysis: quality of care. *Crit Care Med* 1989;**12**(2):S210-2.

[40] Wagner D, Draper E and Knaus W. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Development of APACHE III. *Crit Care Med* 1989;**12**(2):S199-203.

[41] Wagner D, Knaus W and Bergner M. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Statistical methods. *Crit Care Med* 1989;**12**(2):S194-8.

[42] Draper E, Wagner D, Russo M, Bergner M, Shortell S, Rousseau D *et al*. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Study design--data collection. *Crit Care Med* 1989;**17**(12):S186-93.

[43] Cho DY and Wang YC. Comparison of the APACHE III, APACHE II and Glasgow Coma Scale in acute head injury for prediction of mortality and functional outcome. *Intensive Care Med* 1997;**23**(1):77-84.

[44] Castella X, Artigas A, Bion J and Kari A. A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. The European/North American Severity Study Group. *Crit Care Med* 1995;**23**(8):1327-35.

[45] Bastos PG, Sun X, Wagner DP, Knaus WA and Zimmerman JE. Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. *Intensive Care Med* 1996;**22**(6):564-570.

[46] Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C and Kanus WA. Evaluation of Acute Physiology and Chronic Health Evaluation III predictions of hospital mortality in an independent database. *Crit Care Med* 1998;**26**:1317-1326.

[47] Rivera-Fernandez R, Vazquez-Mata G, Bravo M, Aguayo-Hoyos E, Zimmerman J, Wagner D *et al*. The Apache III prognostic system: customized mortality predictions for Spanish ICU patients. *Intensive Care Med* 1998;**24**(6):574-81.

[48] Hosmer DW, Hosmer T, Le Cessie S and Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;**16**(9):965-980.

[49] Lemeshow S, Teres D, Pastides H, Avrunin JS and Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. *Crit Care Med* 1985;**13**(7):519-525.

[50] Lemeshow S, Teres D, Avrunin JS and Gage RW. Refining intensive care unit outcome prediction by using changing probabilities of mortality. *Crit Care Med* 1988;16(5):470-477.

[51] Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH and Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993;**270**(20):2478-2486.

[52] Lemeshow S, Klar J, Teres D, Avrunin JS, Gehlbach SH and Rapoport J. Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective, multicenter study. *Crit Care Med* 1994;22(9):1351-1358.

[53] Moreno R, Miranda DR, Fidler V and Van Schilfgaarde R. Evaluation of two outcome prediction models on an independent database. *Crit Care Med* 1998;**26**(1):50-61.

[54] Le Gall JR, Loirat P, Alperovitch A, Glaser P, Pranthil C, Mathieu D *et al*.. A simplified acute physiology score for ICU patients. *Crit Care Med* 1984;**12**:975-977.

[55] Le Gall JR, Lemeshow S and Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993;**270**(24):2957-2963.

[56] Auriant I, Vinatier I, Thaler F, Tourneur M and Loirat P. Simplified acute physiology score II for measuring severity of illness in intermediate care units. *Crit Care Med* 1998;**26**(8):1368-71.

[57] Sleigh JW, Brook RJ and Miller M. Time-dependent error in the APACHE II scoring system. *Anaesth Intensive Care* 1992;**20**(1):63-5.

[58] Wagner DP, Knaus WA, Harrell FE, Zimmerman JE and Watts C. Daily prognostic estimates for critically ill adults in intensive care units: Results from a prospective, multicenter, inception cohort analysis. *Crit Care Med* 1994;**22**(9):1359-1372.

[59] Sicignano A, Carozzi C, Giudici D, Merli G, Arlati S and Pulici M. The influence of length of stay in the ICU on power of discrimination of a multipurpose severity score (SAPS). *Intensive Care Med* 1996;**22**(10):1048-51.

[60] Rumelhart DE, Hinton GE and Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;**323**(6188):533-536

[61] Levin E, Gewirtzman R and Inbar GF. Neural network architecture for adaptive system modeling and control. *Neural Networks* 1991;**4**(2):185-191.

[62] De Laurentiis M and Ravdin PM. Survival analysis of censored data: Neural network analysis of complex interactions between variables. *Breast Cancer Research and Treatment* 1994;**32**:113-118.

[63] Faraggi D and Simon R. A neural network model for survival data. *Stat Med* 1995;**14**:73-82.

[64] Patil S, Henry JW, Rubenfire M and Stein PD. Neural network in the clinical diagnosis of acute pulmonary embolism. *Chest* 1993;**104**:1685-89.

[65] Duh M-S, Walker AM, Pagano M and Kronlund K. Prediction and cross-validation of neural networks versus logistic regression: Using hepatic disorders as an example. *Am J Epidemiol* 1998;**147**:407-13.

[66] Lapuerta P, Rajan S and Bonacini M. Neural networks as predictors of outcomes in alcoholic patients with severe liver disease. *Hepatology* 1997;**25**:302-306.

[67] Maddrey WC, Boitnott JK, Bedine MS, Weber FL, Mezey E and White RI. Corticosteroid therapy of alcoholic hepatitis. *Gastroenterology* 1978;**75**:193-9.

[68] Heden B, Ohlsson M, Edenbrandt L, Rittner R, Pahlm O and Peterson C. Artificial neural networks for recognition of electrocardiographic lead reversal. *Am J Cardiol* 1995;**75**:929-33.

[69] Lapuerta P, Anzen SP, and LaBree L. Use of neural networks in predicting the risk of coronary artery disease. *Computers and Biomedical Research* 1995;**28**:38-52.

[70] Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. *Ann Intern Med* 1991;**115**:843-48.

[71] Baxt WG. Analysis of the clinical variables driving decision in an artificial neural network trained to identify the presence of myocardial infarction. *Ann Emerg Med* 1992;**21** :1439-44.

[72] Mobley BA, Leasure R and Davidson L. Artificial neural network predictions of lengths of stay on a post-coronary care unit. *Heart and Lung* 1995;**24**:251-6.

[73] Tu JV and Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. *Proc Annu Symp Comput Appl Med Care* 1992; 666-672.

[74] Tu JV, Weinstein MC, McNeil BJ and Naylor CD. Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? *Med Decis Making* 1998;**18**(2):229-35

[75] Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996;**49**:1225-31.

[76] Lane PL, Doig GS, Stewart TC, Mikrogianakis A and Stefanits T. Trauma outcome analysis and the development of regional norms. *Accident Analysis and Prevention* 1997;**29**(1):53-56.

[77] Rutledge R. Injury severity and probability of survival assessment in trauma patients using a predictive hierarchical network model derived from ICD-9 codes. *The Journal of Trauma: Injury, Infection and Critical Care* 1995;**38**(4):590-601.

[78] McGonical MD, Cole J, Schwab CW, Kauder DR, Rotondo MF and Angood PB. A new approach to probability of survival scoring for trauma quality assurance. *J Trauma* 1993;**34**(6):863-70.

[79] Dybowski R, Weller P, Chang R and Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *Lancet* 1996;**347**:1146-50.

[80] Doig GS, Inman KJ, Sibbald WJ, Martin CM and Robertson JMcD. Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. *Proc Annu Symp Comput Appl Med Care* 1993;361-365.

[81] SAS® Institute Inc., *SAS® User's Guide: Statistics, Version 6, 4th Edition*, Cary, NC: SAS® Institute Inc. 1990.

[82] Hosmer DW and Lemeshow S. *Applied Logistic Regression*. New York: John Wiley and Sons, 1989.

[83] Belsley DA, Kuh E and Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: John Wiley and Sons, 1980.

[84] Kleinbaum DG, Kupper LL, Muller KE and Nizam A. *Applied Regression Analysis and Other Multivariable Methods, 3rd Edition*, Pacific Grove, California: Duxbury Press, 1998.

[85] Hosmer DW and Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. *Commu in Stat* 1980;**A10**:1043-1069.

[86] Harrell FE, Lee KL, Califf RM, Pryor DB and Rosati RA. Regression modeling strategies for improved prognostic prediction. *Stat Med* 1984;**3**(2):143-152.

[87] Iezzoni LI. *Risk Adjustment for Measuring Health Care Outcomes.* Ann Arbor, Michigan: Health Administration Press, 1994.

[88] Ward Systems Group, Inc., Executive Park West, 5 Hillcrest Dr., Frederick, MD. 21703, U.S.A.

[89] Beeler G. *Personal communications*. Area chair, *Proc Annu Symp Comput Appl Med Care* 1993.

[90] Roe CA and Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. *Acad Radiol* 1997;**4**(4):298-303.

[91] Metz CE, Wang P-L and Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. *In: Information Processing in Medical Imaging* (F. Deconinck, ed.). The Hague: Martinus Nijhoff, 1984.

[92] Kleinbaum DG, Kupper LL and Morgenstern H. *Epidemiologic Research: Principles and quantitative methods*. New York: Van Nostrand Reinhold, 1982.

[93] Australian and New Zealand Faculty of Intensive Care Policy Documents. Minimum standards for intensive care units. Review IC-1. 1997.

[94] Donabedian A. *Explorations in quality assessment and monitoring. Volume 1. The definition of quality and approaches to its assessment.* Ann Arbor, MI: Health Administration Press, 1980.

[95] Fink MP, Snydman DR, Niederman MS, Leeper KV Jr, Johnson RH, Heard SO, *et al*. Treatment of severe pneumonia in hospitalized patients: results of a multicenter, randomized, double-blind trial comparing intravenous ciprofloxacin with imipenem-cilastatin. The Severe Pneumonia Study Group. *Antimicrob Agents Chemother* 1994;**38**(3):547-557.

[96] Hebert PC, Wells G, Blajchman MA, Marshall J, Martin CM, Pagliarello G, *et al.* A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. *N Engl J Med* 1999;**340**:409-17.

[97] Harrell FE, Lee KL and Mark DB. Tutorial in Biostatistics: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361-387.

[98] Tu JV and Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. *Comp Biomed Res* 1993;**26**:220-229.

[99] Breslow NE, Day NE, Halvorsen KT, Prentice RL and Sabai C. Estimation of multiple relative risk functions in matched case-control studies. *Am J Epidemiol* 1978;**108**(4):299-307.

[100] Lemeshow S and Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol* 1982;**115**(1):92-106.

[101] Hosmer DW, Taber S and Lemeshow S. The importance of assessing the fit of logistic regression models: A case study. *Am J Public Health* 1991;**81**:1630-1635.

[102] Kattan MW and Beck JR. Artificial neural networks for medical classification decisions. *Arch Pathol Lab Med* 1995;**119**:672-677.

[103] Bright P, Miller MR, Franklyn JA and Sheppard MC. The use of a neural network to detect upper airway obstruction caused by goiter. *Am J Respir Crit Care Med* 1998;**157**:1885-91.

[104] Buchman TG, Kubos KL, Seidler AJ, Siegforth MJ. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit. *Crit Care Med* 1994;**22**:750-762.

[105] Kung SY. *Digital Neural Networks*. Englewood Cliffs, New Jersey: PTR Prentice-Hall Inc, 1993.

[106] Jennett B, Teasdale G, Braakman R, Minderhoud J and Knill-Jones R. Predicting outcome in individual patients after severe head injury. *Lancet* 1976;**1**(7968):1031-4.

[107] Bone RC. Sir Isaac Newton, sepsis, SIRS, and CARS. *Crit Care Med* 1996;**24**(7):1125-1128.

[108] Hornik K, Stinchcombe M and White H. Multilayer feedforward networks are universal approximators. *Neural Networks* 1989;**2**:359-366.

[109] Lippmann RP. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 1987;**4**:4-22.

[110] Knaus WA, Harrell FE, Lynn J, Goldman L, Philips RS, Connors AF *et al*. The SUPPORT Prognostic Model: Objective estimates of survival for seriously ill hospitalized adults. *Ann Intern Med* 1995;**122**:191-203.

[111] Hosmer DW and Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine* 1995;**14**:2161-2172.

[112] Teres D and Lemeshow S. Why severity models should be used with caution. *Crit Care Clinics* 1994;**10**(1):93-110.

[113] Rapoport J, Teres D, Lemeshow S and Gehlbach S. A method for assessing the clinical performance and cost-effectiveness of intensive care units: a multicenter inception cohort study. *Crit Care Med* 1994;**22**(9):1385-1391.

Appendix I

Back-propagation artificial neural networks:

A primer.

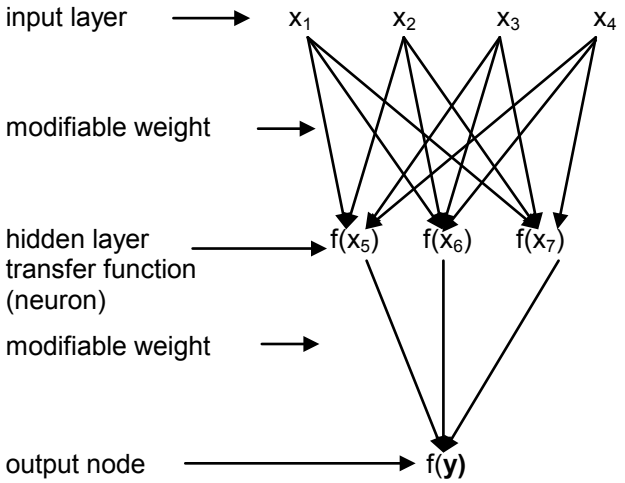
The methods by which the human mind processes information and learns to solve new problems are still largely a mystery. In order to understand these processes, cognitive science researchers have developed many different types of artificial ‘thinking machines’. Within the last few years, some interesting results have been achieved by making the architecture of these thinking machines mimic the physical structure of the human mind.

In the human brain, the actual information processing is performed by a very simple element: the neuron. A typical neuron receives incoming signals from other neurons through a host of fine input structures called dendrites. The neuron processes the incoming signal and then sends out spikes of electrical activity through its axon, which splits into thousands of branches that connect with the dendrites of other neurons. ANNs are also composed of information processing 'neurons' that are interconnected to other similar neurons. The resulting networks of artificial neurons are simulated in a digital computer and are nowhere near as complex as the neural networks in the human mind that they are supposed to represent.

The topology, or layout, of the network determines the functions that it is able to perform. The presence of interconnections between neurons determines whether it is possible for one neuron to influence another. By setting up different patterns of interconnections between the artificial neurons, researchers have been able to rule out all kinds of theories about how the human brain processes information and have gained remarkable insights into how the mind actually learns. One interesting finding is that it is not the inherent information processing abilities of the human neuron that lends us our ability to solve complex problems, but rather the complexity of the interconnections between these neurons.

Biological neural networks are typically arranged in layers. For artificial networks, the most common arrangement is to have three layers of neurons; an input layer which passes information to a middle or hidden layer which in turn passes information to a final output layer. The sole function of the input layer is to distribute the raw information that is fed into the network to the hidden layer. The hidden layer passes the information to the output layer, which produces results that can be interpreted by the user. The results produced by the output layer can be continuous, dichotomous or even categorical.

Simplified Neural Network



The most important network function is performed by the hidden layer. It is the neurons in this layer that have the job of associating a particular input pattern with the appropriate desired output values. The ability of the hidden layer to perform this job surprisingly well has made neural networks such useful tools.

The activity level, or value, of each hidden unit is determined by the activity of the input units and the strength of the connections between the input and hidden units. Similarly, the value of the output unit depends on the activity of the hidden units and the strength of the connections between the hidden and output units. The strength of the connections between neurons is determined by a numerical value or weight given to the interconnection. By being allowed to modify these weights, a hidden unit can learn to associate different inputs in such a way that the desired value of output is achieved.

Training a network to perform a particular task is achieved by presenting the network with example cases. Each example case consists of a group (vector) of predictor

variables presented to the input units together with the desired value of the outcome, which is presented to the output units. By constraining the input and the output layers, the hidden layer is forced to adapt to find a solution. The hidden layer uses a systematic algorithm to change the interconnection weights between itself and the input layer and between itself and the output layer such that the input values result in a final output as close as possible to the desired outcome value. In effect, the network searches for the appropriate transformation to map the predictor values onto the desired outcome. It starts with a transformation function chosen by the user based on the type of problem to be solved, and through a process of iterative error reduction, optimizes the parameters associated with this transformation until an acceptable solution is found. The parameters associated with the transformation are represented by the network's interconnection weights.

Obviously the approach used to change the interconnection weights is much more complex than described above. In fact, in order to calculate the appropriate amount to change each weight we must first calculate the rate at which the total output error changes in response to the proposed weight change. First year calculus tells us that the first order derivative of any function represents the rate of change of that function. This means that the rate of change of the overall output error is represented by the first order derivative of the total output error function. This quantity is referred to as the error derivative for the weight (EW) and must be calculated for each connection in order to determine how much its weight should be adjusted.

The most intuitive method used to find the EW is a simple iterative 'trial and error' process. Each weight on the network is adjusted after each case, very slightly, and one at a time. The effect of each weight change on the total output error is observed. In all but the most simple networks, this method is computationally inefficient, even for today's most powerful workstations. Although there are many possible approaches to the calculation of the EW, the method used by most networks today is called back-propagation of the errors. Technical details of this algorithm are provided in the next section of this appendix.

Once the EW is derived for each connection, the appropriate amount of change required for a particular weight can be calculated. It should be kept in mind that each weight can only be changed very slightly so as not to start increasing the error in the opposite direction. The total output error is measured by summing the square of the differences between all of the actual outputs and the expected outputs, which are provided by the training set. This process of weight changes is repeated until the total error reaches a minimum.

For example, if we assume that we have 100 training cases, the network weights would be adjusted after all the cases had been presented at least once. As long as the weight adjustments result in a decrease in the total error, we will keep presenting the same 100 cases to the network until a minimum error is reached. The same 100 cases can be presented to the network 100 times or more. Learning is terminated automatically when the total output error ceases to decrease in response to the presentation of the test cases.

Using the back-propagation algorithm, multi-layered networks have been taught to play backgammon, predict the secondary structure of proteins and recognize precancerous Pap-smears. Networks have also been trained to recognize hand written letters, interpret ECG strips and predict currency exchange rates.

Many topologies or layouts are possible for ANNs as there are many different approaches to finding the error derivative of the weights. The reason that back-propagation networks have generated unprecedented interest by statisticians and other primary researchers is that they have the ability to simulate many statistical functions surprisingly well. Simple single-layered back-propagation ANNs can simulate general linear and logistic regression and three-layered networks can simulate multiple regression that considers all possible interactions between the predictor variables. Neural networks with multiple output nodes can even perform regression on polychotomous outcomes. Unfortunately, the main drawback of neural network techniques is their inability to estimate the overall effect of individual input parameters. They have, however, proven to be extremely accurate at predicting outcomes.

Back-propagation neural networks can perform complex time series predictions. Researchers have shown that multi-layered back-propagation networks can perform very accurate predictions on time series generated by nonlinear delayed differential equations. The predictive error of these networks was lower than complicated techniques such as linear predictive methods, the local linear method of Farmer and Sidorowich, and direct and iterative Gabor polynomial approaches.

The most exciting application for the predictive powers of back-propagation networks comes from the engineering discipline of systems analysis, especially with respect to adaptive control systems. This field deals with the control aspects of complex feedback systems, which are inherently difficult to predict since their behavior is usually defined by both deterministic and stochastic processes. The stochastic nature of these systems makes it extremely hard to distinguish between noise in the input parameters and irregularities due to deterministic but nonlinear dynamic behavior. Linear prediction fails miserably due to the complex behavior patterns exhibited by such systems. Construction of the optimal nonlinear prediction function requires the *a priori* knowledge of the properties of the underlying nonlinear comprehensive equations for the system. Although no method can predict the stochastic component of the systems behavior, neural networks actually provide optimum solutions under most situations, as compared directly to the comprehensive defining equations. When these equations are not known *a priori*, the networks provide the only useful method of optimal prediction of the future responses to changes in the input parameters.

Recent advances in the understanding of the nature of cardiac physiology and immunology, have shown the potential of these systems to display complex nonlinear behavior. In response to certain inputs, such as exaggerated internal feedback loops, massive disease or aggressive treatment, many of the body's biological systems can exhibit strange and unexpected responses. These types of responses are typical of the more complex nonlinear dynamic adaptive control systems dealt with by engineers. If the underlying comprehensive equations were known, then these strange nonlinear responses could be predicted and avoided. Since ANNs have the ability to predict the responses of nonlinear dynamic adaptive control systems optimally and they can simulate complex

statistical operations, perhaps it is time to investigate thoroughly their potential to deal with problems faced in such multidisciplinary areas as intensive care medicine.

The Back Propagation Algorithm

In order to reduce the error between the actual and desired output, the neural network must compute the error derivative of the modifiable weights (*EW*) for each connection to and from the hidden layer. The EW measures the rate of change of the error in response to a small change of a single interconnection weight. The back-propagation algorithm is the most widely used method for determining the *EW*.

In order to understand the back-propagation algorithm, the neural network must first be described in mathematical terms. Assume that unit *j* is a typical unit in the output layer and unit *i* is a typical unit in the hidden layer. The output unit determines its activity level by following a two-step procedure. First, it computes the total input from the hidden layer (x_j,) using the formula

$$x_j = \sum_i y_i w_{ij} ,$$

where y_i is the activity level of the i^{th} unit in the previous layer and w_{ij} is the weight of the connection between the i^{th} and j^{th} unit. In other words, the output unit simply sums all the incoming signals from the hidden layer.

Next, the output unit passes this summed input through its own simple internal processing function. This results in the actual output y_j. In most back-propagation networks, the logistic function is used:

$$y_j = \frac{1}{1 + e^{-x_j}}.$$

Once the activity of the output unit has been determined, the network computes the total error E, which is defined by the expression;

$$\mathrm{E} = \frac{1}{2}\sum_j (y_j - d_j)^2,$$

where y_j is the activity level of the output unit for the j^{th} case and d_j is the desired output of the j^{th} case. Put simply, this is one half of the sum of the squares of the differences

between the actual and desired outputs. It is this quantity that we wish to minimize through iterative applications of the back-propagation algorithm.

The back-propagation algorithm consists of four steps:

1. Compute how fast the error changes (rate of change = first derivative of error function) as the activity of an output unit is changed. This error derivative (*EA*) is the difference between the actual and the desired activity

$$EA_j = \frac{\delta \mathrm{E}}{\delta y_j} = y_j - d_j$$

2. Compute how fast the error changes as the *total* input received by the output unit is changed. This quantity (*EI*) is the answer from step 1 multiplied by the rate at which the output of a unit changes as its total input is changed.

$$EI_j = \frac{\delta \mathrm{E}}{\delta x_i} = \frac{\delta \mathrm{E}}{\delta y_j}\frac{dy_j}{dx_j} = EA_j y_j (1 - y_j)$$

3. Compute how fast the error changes as a single weight on a single connection into the output unit is changed. This quantity (*EW*) is the answer from step 2 multiplied by the activity level of the unit from which the connection emanates.

$$EW_{ij} = \frac{\delta \mathrm{E}}{\delta w_{ij}} = \frac{\delta \mathrm{E}}{\delta x_j}\frac{\delta x_j}{\delta w_{ij}} = EI_j y_i$$

4. Now we have isolated the first *EW* of one single connection, which tells us the effect of changing that connection's weight on the total output. Next we move back up the connection to the node from which this *EW* was calculated. We then compute how fast the error changes as the activity of this new unit in the hidden layer is changed. This crucial step allows back-propagation to be applied to multi-layered networks (and hence the name, we propagate the errors back-up the network). When the activity of a unit in the hidden layer changes, it affects the activity of the output unit to which it is connected. So to compute the overall effect on the error, we add together all these separate effects on the output unit. Each effect is simple to calculate. It is the answer in step 2 multiplied by the weight on the connection to that output unit.

$$EA_i = \frac{\delta \mathrm{E}}{\delta y_i} = \sum_j \frac{\delta \mathrm{E}}{\delta x_j}\frac{\delta x_j}{\delta y_i} = \sum_j EI_j w_{ij}$$

By using steps 2 and 4, we can convert the *EA*s of one layer of units into *EA*s for the previous layer. This procedure can be repeated to get the *EA*s for as many previous layers as desired. Once we know the *EA* of a unit, we can use steps 2 and 3 to compute the *EW*s on its incoming connections. Once we have calculated the *EW* for each connection weight, which serves as a measure of the strength of effect that each weight has on the outcome, each weight is modified in order to reduce the total output error. This weight correction is verified by passing the test cases through the network a second time, and if necessary correcting the weights again. This iterative application of the back-propagation algorithm continues until the magnitude of the total error function stops decreasing.

This brief description of the back-propagation algorithm was adapted from Hinton (1992). The back-propagation technique is one method for finding minima of a nonlinear function *g(x)* in a space of high dimension. For a comprehensive development, which is beyond the scope of this appendix, refer to Müller and Reinhardt (1991).

Further Readings on Neural Networks

Caudill M and Butler M. *Naturally intelligent systems*. The MIT Press, 1991.

Hertz J, Krogh A and Palmer RG (eds). *Introduction to the theory of neural computation*. Addison-Wesley, 1990.

Hinton GE. Connectionist learning procedures. *Artificial Intelligence* 1989;**40**;185-234.

Hinton GE. How neural networks learn from experience. *Sci Am* 1992;**267**:145-151.

McClelland JL, Rumelhart DE and the PDP Research Group. *Parallel distributed processing, explorations in the microstructure of cognition, Volume 1: Foundations*. The MIT Press, 1987.

Müller B and Reinhardt J. *Physics of neural networks - an introduction*. Springer-Verlag, New York, 1991.

Rosenberg CR and Sejnowski TJ. Parallel networks that learn to pronounce English text. *Complex Systems* 1987;**1**:145-168.

Rumelhart DE, Hinton GE and Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;**323**(6188):533-536.

Appendix II

Copyright release from the

American Medical Informatics Association

Appendix III

Modeling mortality in the intensive care unit:

A pilot project

Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression.

Gordon S. Doig[1], Kevin J. Inman[2], William J. Sibbald[2], Claudio M. Martin[2], and James McD. Robertson[1]

[1]Department of Epidemiology and Biostatistics, University of Western Ontario,

[2]The Richard Ivey Critical Care Trauma Centre, Victoria Hospital Corp., London, Ontario, Canada

ABSTRACT

The objective of this study was to compare and contrast two techniques of modeling mortality in a 30 bed multi-disciplinary ICU; neural networks and logistic regression. Fifteen physiological variables were recorded on day 3 for 422 consecutive patients whose duration of stay was over 72 hours. Two separate models were built using each technique. First, logistic and neural network models were constructed on the complete 422 patient dataset and discrimination was compared. Second, the database was randomly divided into a 284 patient developmental dataset and a 138 patient validation dataset. The developmental dataset was used to construct logistic and neural net models and the predictive power of these models was verified on the validation dataset. On the complete dataset, the neural network clearly outperformed the logistic model (sensitivity and specificity of 1 and .997 vs. .525 and .966, area under ROC curve .9993 vs. .9259), while both performed equally well on the validation dataset (area under ROC of .82). The excellent performance of the neural net on the complete dataset reveals that the problem is classifiable. Since our dataset only contained 40 mortality events, it is highly likely that the validation dataset was not representative of the developmental dataset, which led to a decreased predictive performance by both the neural net and the logistic regression models. Theoretically, given an extensive dataset, the neural network should be able to perform mortality prediction with a sensitivity and a specificity approaching 95%. Clinically, this would be an extremely important achievement. In future trials, we intend to investigate the performance of an application-specific, state of the art neural network on a more representative, comprehensive prospective patient database.

ROC = receiver operating characteristic

INTRODUCTION

Today, it is common practice to assign a severity-of-illness score to a patient upon entry into the intensive care unit (ICU). Common ICU scoring systems include the Acute Physiology and Chronic Health Evaluation (APACHE), the Mortality Probability Models (MPMs), the Simplified Acute Physiology Score (SAPS) and the Pediatric Risk of Mortality (PRISM) scoring system [1].

The scoring strategies for each of these systems differ markedly, but all combine measures of current physiological status with various preexisting risk factors to produce a surrogate measure of risk or disease severity. This surrogate measure can then be used as a tool to aid in quality assurance, resource allocation, clinical decision making, the evaluation of new therapies and outcome prediction [1,2,3].

In the APACHE and MPM approach to predicting outcomes, the day 1 risk factors identified by each model are entered into a logistic regression equation which then produces a predicted probability of mortality. These logistic regression models usually perform very well when predicting the expected mortality experience of an ICU but fall short of clinical usefulness when predicting mortality of the individual patient [2,4].

Recent research has shown that events occurring after ICU admission are more useful predictors of outcomes than ICU admission status [4]. It has also been shown that if the APACHE II scores remain high in the face of continued maximal intervention, fatal

outcome can be predicted [5]. These studies indicate that a scoring system enacted some time after admission should have a better predictive performance than a scoring system enacted at admission.

Some researchers believe that understanding the patterns that are associated with survival or death may require the use of alternative mathematical approaches, such as set theory or fuzzy logic, which may ultimately be more fruitful than further attempts at refining existing systems [5, 6].

Alternative techniques such as nonlinear discrete neural networks, have recently begun to be applied to some classical medical problems. These techniques are derived from the engineering disciplines of pattern recognition and signals processing and are extremely promising because they offer the potential for ever-improving performance through dynamic learning [7].

A neural network trained on 351 patients with a high likelihood of myocardial infarction outperformed emergency room physicians when presented with 331 new cases of patients presenting with anterior chest pain. The physicians diagnosed myocardial infarction with a sensitivity and specificity of 77% and 84% respectively, whereas the neural network performed with a sensitivity and specificity of 97.2% and 96.2% [8].

Neural networks have outperformed clinicians on the diagnosis of hepatic masses [9], pulmonary emboli [10], and breast tumors [11]. Artificial neural networks have also shown their potential usefulness in the ICU by predicting the length of ICU stays after cardiac surgery [12].

The purpose of this study was to compare and contrast the performance of a relatively simple back-propagation, associative-learning neural network with a classical multivariate logistic regression approach to predicting ICU mortality based on day 3 physiology scores.

METHODS

Patient Selection

During a six month period from August 5, 1991 to February 5, 1992, 614 patients were admitted to the 30 bed multi-disciplinary adult critical care unit. The only entry criterion for this study was a duration of stay greater than 72 hours. Four hundred and twenty-two patients met this criterion and were therefore eligible.

Data Collection

Fifteen variables were recorded daily for the duration of stay for each study entrant. These variables were identified from the literature [13] and from clinical experience. They were; presence of acute renal failure, packed cell volume, heart rate, FIO_2, serum sodium, PaO_2, pH, respiratory rate, systolic and diastolic blood pressures, serum potassium, temperature, white blood cell count, serum creatinine and the Glasgow coma score. The definition and recording of all variables was consistent with the methods outlined for data collection for the APACHE II scoring system [13].

The outcome of interest was ICU mortality in patients with a duration of stay greater than 72 hours.

During the period of the study, the variables were abstracted from patient records and stored in a central codebook. At the conclusion of the study, the codebook was entered into a spreadsheet program and then transferred into PC SAS® version 6.04[1].

Database Validation

Primary data integrity was verified in PC SAS® with algorithms written to filter out biological impossibilities and obvious data entry transpositions. Any values in conflict with the screening filters were re-entered directly from the study codebook.

Secondary validation was carried out by match-merging the study database with a readily available XENIX based ICU management information system (MIS). This MIS database allowed validation of date of birth, ICU entry date, ICU discharge date and ICU discharge status. Since the MIS database is utilized for billing purposes, its entries are double-verified and seldom in error. Conflicts with the study database were resolved by accepting the MIS database as correct.

Primary and secondary error detection rates were then compared as a means to increase confidence in data integrity.

Scale Selection

Variables such as heart rate can convey different information about clinical interventions, outcomes and risks dependent upon the degree of elevation or depression above or below the normal range. For this reason, all variables except serum creatinine, presence of acute renal failure and the Glasgow coma score were separated into high or low distributions about the median. This resulted in a total of 27 variables.

In order to present the variables to the neural network, scaling was required. Each high or low distribution variable was non-parametrically transformed to the z-scale. All 27 z-transformed variables were presented to the neural network and the logistic model to maintain consistency.

Logistic Regression

Logistic regression was performed using PROC LOGIST, PC SAS®, version 6.04 [14]. Variables were considered as candidates for inclusion in the model based on a univariate logistic regression p-value 7 0.25 [15]. A multiple-step backward model selection method was used and variables were removed from the model if significance fell above a p-value of 0.10. After the final model was evaluated, first order interaction terms were assessed.

Neural Network

A commercially available back-propagation, associative-learning neural network was used for this simulation[2]. All 27 variables were presented to a 3 layered network with 27 input nodes, 18 hidden nodes and 1 output node. A logistic activation function was used and the output node generated a probability of mortality ranging between 0 and 1.

---

[1]PC SAS® version 6.04, SAS Institute Inc., SAS Circle, PO Box 8000,Cary, NC, 27512-8000, U.S.A.

[2]NeuroShell™,Ward Systems group, Inc., 245 W. Patrick St., Frederick, MD 21701, U.S.A.

Through an error-minimization technique known as back-propagation, the neural network optimizes weights between nodes such that important patterns between variables are recognized.

Comparisons

The performance of the neural network and the logistic model were compared under two different conditions. First the neural network and the logistic regression techniques were exposed to the complete database. Their ability to discriminate between patients who lived or died was then compared.

Second, a developmental and validation subset were randomly selected from the complete dataset. The developmental dataset contained 284 patients and was used to create a new logistic model and a new neural network model. The two models were then rated on their ability to discriminate between patients who lived or died in the validation dataset, to which they had never been exposed.

Discrimination was assessed using the area under the receiver operating characteristic (ROC) curve of each model [16]. Performance was also assessed by comparing the sensitivity and specificity of the approaches at the arbitrary classification threshold of 0.5.

RESULTS

Patient Population

The average age of the study subjects was 61 years and the average duration of ICU stay was 7.3 days. For the period of the study, the average day 1 APACHE II score was 25.3. The study population experienced a 9.5% mortality rate.

Database Validation

Primary validation revealed a coding error rate of 2.1 % and independent secondary validation against the MIS database revealed an error rate of 2.4%. Comparison of study database errors against codebook values did not reveal any obvious transposition errors in the codebook. Only 0.2% of values were missing due to initial failure to perform laboratory tests at bedside.

Logistic Regression

Complete Dataset

The final model contained seven significant physiological variables. They were; presence of acute renal failure (ARF), high serum sodium, high pH, high diastolic blood pressure (DBP), the Glasgow coma score (GCS), high $PaCO_2$, and low serum sodium (see table 1).

Table 1	β	SE	OR	p
Intercept	-1.6136	0.98	0.199	0.101
ARF	2.3518	0.59	10.50	0.000
Hi Na	0.9741	0.28	2.648	0.000
Hi pH	0.9659	0.39	2.627	0.014
Hi DBP	-1.1326	0.38	0.322	0.003
GCS	-0.2390	0.06	0.787	0.000
Hi $PaCO_2$	-0.8835	0.49	0.413	0.076
Low Na	1.6940	0.40	5.441	0.000

β=regression parameter estimate, SE=standard error of regression parameter, OR=odds ratio, p= p-value

At a classification threshold of 0.5, the logistic regression model performed with a sensitivity of .525 and a specificity of .966. The positive predictive value was .618 and the negative predictive value was .951. The area under the ROC curve was .9259.

Validation Dataset

The developmental dataset of 284 patients produced a final model containing seven different significant variables. The seven variables were; high serum sodium, high diastolic blood pressure, GCS, high PaCO2, low serum sodium, low serum potassium, and low temperature (see table 2).

At a classification threshold of 0.5, this model performed with a sensitivity of .133 and a specificity of .976. The positive predictive value was .400 and the negative predictive value was .902. The area under the ROC curve was .8320.

Table 2	β	SE	OR	p
Hi Na	1.2876	0.34	3.624	0.000
Hi DBP	-1.8092	0.47	0.164	0.000
GCS	-0.2357	0.06	0.790	0.000
Hi PaCO2	-1.0222	0.50	0.360	0.042
Low Na	1.8762	0.64	6.529	0.003
Low K	2.0746	0.80	7.961	0.009
Low Temp	0.4771	0.22	1.611	0.029

Neural Network

Complete Dataset

The network converged on a solution after 15,837,750 iterations. This took 17:10:43 hours on a 27 MHz 386.

At a classification threshold of 0.5, the neural network performed with a sensitivity of 1.0 and a specificity of .997. The positive predictive value was .976 and the negative predictive value was 1.0. The area under the ROC curve was .9993.

Validation Dataset

Using the 284 patient developmental database, the network converged on the optimum predictive solution after 20,300 iterations, which took 17:21 minutes on a 27 MHz 386.

At a classification threshold of 0.5, the neural network classified the 138 patient validation database with a sensitivity of .267 and a specificity of .976. The positive predictive value was .571 and the negative predictive value was .916. The area under the ROC curve was .8178.

DISCUSSION

Back-propagation neural networks have traditionally excelled at classification (pattern recognition) problems. They are most useful in situations where the relationship between the input and the output is nonlinear and training data are abundant [17].

On the complete dataset, the back-propagation network clearly outperforms logistic regression with respect to the classification of mortality and survivability (sensitivity and specificity of 1.0 and .997 verses .525 and .966).

With the neural network performing with an area under the ROC curve of .9993 and only one misclassified event, we can conclude that the 15 recorded day 3 physiological variables adequately describe the mortality patterns experienced over the period of the study.

Baxt's neural network predicted myocardial infarction by placing diagnostic importance on clinical variables that have not previously been shown to be highly predictive for infarction [18]. Since the etiology of mortality is much more complex than the etiology of infarction, and since discrimination was so successful with our neural

network, it suggests that patterns and predictors of mortality are being detected that were not detected using the traditional logistic regression approach.

When 2/3 of the complete dataset was used for model building and 1/3 for validation, the overall predictive performance of the two approaches was identical (area under ROC = .82). The neural network was more sensitive over the range of decision thresholds while the logistic model was more specific.

Performance of any predictive model on a validation dataset depends on how representative the validation cases are of the developmental cases and on how well the model can classify the developmental dataset. Theoretically, if the validation dataset is truly representative of the developmental dataset, then the predictive performance will approach the level of developmental classification.

If the neural network were exposed to more cases, in the form of a larger dataset, there is no reason to suspect that a similar level of classification would not occur. If this dataset were extensive enough to cover most patterns of mortality, then a predictive sensitivity and specificity of over 95% could reasonably be expected.

Clinically this would be an extremely important achievement. Improved predictive performance would enhance quality assurance, resource allocation, and the evaluation of new therapies. With a sufficiently high predictive performance, the neural network would also be an unprecedented ancillary aid in clinical decision making at the individual level.

Primary research in neural networks is a dynamic and rapidly progressive field. We intend to investigate the performance of an application-specific, state of the art neural network on a more representative, comprehensive prospective patient database.

References

[1]. Seneff M, Knaus WA. Predicting patient outcome from intensive care: a guide to APACHE, MPM, SAPS, PRISM and other prognostic scoring systems. J Intensive Care Med 1990;5:33-552

[2]. Rutledge R, Fakhry SM, Rutherford EJ, Muakkassa F, Baker CC, Koruda M and Meyer AA. Acute Physiology and Chronic Health Evaluation (APACHE II) score and outcome in the surgical intensive care unit: an analysis of multiple intervention and outcome variables in 1,238 patients. Crit Care Med 1991;18:1048-1053

[3]. Knaus WA, Wagner DP, Draper EA, Simmerman JE, Bergner M, Bastos PG, Sirion CA, Murphy DJ, Lotring T, Damiano A and Harrel FE. The APACHE III Prognostic System; Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619-1636.

[4]. Ferraris VA, Propp ME. Outcome in critical care patients: A multivariate study. Crit Care Med 1992;20:967-976.

[5]. Civetta JM, Hudson-Civetta JA, Kirton O, Aragon C and Salas C. Further appraisal of APACHE II limitations and potential. Surg Gynecol Obstet 1992;175:195-203.

[6]. Chang RWS, Jacobs S and Lee B. Predicting outcome among intensive care unit patients using computerized trend analysis of daily Apache II scores corrected for organ system failure. Intensive Care Med 1988;14:558-566.

[7]. Levin E, Gewirtzman R, Inbar GF. Neural network architecture for adaptive system modeling and control. Neural Networks 1991;4:2:185-191.

[8]. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med. 1991;115:843-848.

[9]. Maclin PS and Dempsey J. Using an artificial neural network to diagnose hepatic masses. J Med Syst 1992;16: 215-225.

[10]. Scott JA and Palmer EL. Neural network analysis of ventilation-perfusion lung scans. Radiology 1993;186:661-664.

[11]. Goldberg V, Manduca A, Ewert DL, Gisvold JJ, Greenleaf JF. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. Med Phys. 1992;19:1475-1481.

[12]. Tu JV and Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Proc Annu Symp Comput Appl Med Care 1992; 666-672.

[13]. Knaus WA, Draper EA, Wagner DP, and Zimmerman. APACHE II: A severity of disease classification system. Crit Care Med 1985;13:818-830.

[14]. SAS® Institute Inc., SAS® User's Guide: Statistics, Version 6, 4th Edition, Cary, NC: SAS® Institute Inc. 1990.

[15]. Hosmer DW and Lemeshow S. Applied Logistic Regression. John Wiley and Sons, New York, 1989.

[16]. Erdreich LS and Lee ET. Use of relative operating characteristic analysis in epidemiology. Am J Epidemiol 1981;114:649-662.

[17]. Hinton GE. How neural networks learn from experience. Sci Am 1992;267:145-151.

[18]. Baxt WG. Analysis of the clinical variables driving decision in an artificial neural network trained to identify the presence of myocardial infarction. Ann Emerg Med 1992;21 :1439-1444.

Appendix IV

Data collection forms

Appendix V

Genetic adaptive learning algorithm networks

In 1859, Charles Darwin proposed the three basic principles that he believed determined the survival of a species under his Theory of Evolution: the reproductive cycle; the force of natural selection; and diversity due to variation. Just as artificial intelligence (AI) researchers have used computer simulations to gain insights into how the human mind learns and solves problems, artificial evolution (AE) researchers have used computer simulations to gain a better understanding of the process of evolution.

Based on insights gained from AE simulations, computational intelligence (CI) researchers developed the field of evolutionary computing (EC). In the broadest sense, EC involves the development and application of computer-based problem solving algorithms which incorporate the three basic principles of the Theory of Evolution. Problem solving algorithms which incorporate the principles of EC are called evolutionary algorithms (EA).

An EA is essentially an advanced iterative search algorithm. As the first step in solving a problem, an EA randomly generates a *pool* of potential solutions. The EA ranks the performance of each individual solution in this pool using an objective score. Solutions that perform poorly are discarded (*force of natural selection*) and a set of rules or operators is applied to the remaining solutions in order to generate a new pool of solutions with improved performance (*reproduction*).

The performance of these second-generation solutions are ranked and those performing below the cut-off are discarded. This iterative cycle of *natural selection* and *reproduction* is repeated until a solution is found that satisfies the performance-related stopping criteria. The concept of *diversity* is introduced into the reproductive cycle by the rules and operators that determine how the performance of each successive generation of solutions is improved.

An EA is similar to many other algorithms (least-squares regression, back-propagation) that use iterative estimations to solve a problem. The uniqueness of the EA however, lies in the specific operators that are used to introduce diversity. In 1975, John Holland developed a pair of operators that allowed the genetic principles of *crossover/recombination* and *mutation* to be incorporated into EAs. An EA that contains crossover and mutation operators is referred to as a genetic algorithm (GA).

In genetics, crossover/recombination occurs when two different chromosomes combine and exchange DNA sequences to form two unique new chromosomes whereas a mutation occurs when a DNA sequence within one single chromosome is randomly rearranged. In the context of a GA, where the solution space may be composed of a series of rules generated by an expert system (if **a** then **b**, if **b** then **c**, if **c** then not **f**), a crossover operator could generate novel rules based on combinations of existing rules (if **a** then not **f**) whereas a mutation operator could generate a completely new rule by performing a random operation on one single rule (if **a** then ***d***).

GAs have proven to be particularly successful in searching for solutions to timetable scheduling problems. For example, when a university generates a course timetable, the most desirable solution is one without scheduling conflicts for the professors offering the courses *and* where students are able to schedule their courses according to their degree program requirements. Given that class size determines lecture hall requirements, lecture halls are constrained and must be shared between programs, professors and students can only be in one place at one time, professors often teach more than one course per semester for different programs and most degree programs require students to take more than one course per semester from within their own program and from other programs, the problem can become extremely complex.

The first step to solving this timetable problem with a GA would involve using a very basic rule (e.g. professor conflicts) to generate a *pool* of tentative timetables. Each and every tentative timetable in this pool would then be evaluated for conflicts at the room, professor and program level. Tables with a low number of conflicts would be retained and subjected to crossover and mutation operators.

A crossover operator could be programmed to flag the worst days from the best tables and search for replacements for these entire days from other tables. In this way, a timetable that performed well on Monday, Tuesday and Friday would be combined with a table that performed well on Wednesday and Thursday. A mutation operator could be programmed to detect specific problem areas within days and to randomly reorganize only those days. After all crossover and mutation operators had been performed, the new timetables would be evaluated for conflicts and the process repeated until a solution was

found. Although this process may require several thousand iterations to find the ideal timetable, the amount of computing resources and time required would be significantly less than any other search-based approach.

The United States Navy Center for Applied Research in Artificial Intelligence is currently involved in numerous projects investigating the application of GAs to intelligent systems. GAs have been used to improve the ability of an intelligent autopilot to make F-14 carrier landings under adverse conditions and have also been used to enable an intelligent autonomous underwater vehicle to complete missions under fault scenarios that the original designers had not anticipated.

In the field of engineering, GAs have been employed to optimize the design of cam shapes, to aid in airfoil design and to improve the efficiency of computationally intensive simulations. The field of mobile robotics routinely employs GAs to improve the ability of intelligent devices to handle unforeseen circumstances. More recently, primary artificial neural network (ANN) researchers have explored different applications of the basic principles of GAs to ANN optimization.

NeuroShell 2, Release 3.0[3], which is the software package used to develop ANNs for this project, supports the development of *genetic adaptive learning algorithm networks* (GenNets). Since the back-propagation algorithm is already an iterative estimation process (Appendix I), it requires only minor modifications to incorporate the important concepts of GAs.

The standard back-propagation algorithm begins its search for a solution by randomly generating the weights and coefficients for a single neural network. The GenNet algorithm begins by randomly generating the weights and coefficients for a pool of up to 100 such neural networks.

During one iterative learning cycle, the back-propagation algorithm calculates the errors for each and every case in the developmental data set and back-propagates these errors to adjust the weights and coefficients for each and every connection in the network. After this first iteration, the network is presented with all cases in the developmental data

---

[3] Ward Systems Group, Inc., Executive Park West, 5 Hillcrest Dr., Frederick, MD. 21703, U.S.A.

set again, errors are re-calculated and back-propagated. In this way, the back-propagation algorithm causes the single network to converge upon a solution.

The GenNet algorithm follows a similar cycle except errors are back-propagated for each and every network in the 100 network breeding pool. After each iterative cycle, networks are ranked and the poor performers are discarded. A crossover operator is used to generate novel networks by trading weights between successful networks and a mutation operator is used to randomly adjust weights within networks. After this first iteration, networks in the pool are presented with all cases in the developmental data set again, errors are re-calculated and back-propagated, networks in the pool are ranked and crossover and mutation operators are applied. The GenNet algorithm terminates when one of the networks in the pool satisfies the predefined stopping criteria.

Since the GenNet approach develops and evaluates up to 100 potential solutions, it often takes significantly longer to produce a solution than a simple back-propagation algorithm. Simulation exercises have shown that genetic adaptive approaches to developing neural networks often succeed in situations where single networks have problems converging on a solution. Although the genetic adaptive algorithms have demonstrated advantages over the basic back-propagation algorithm, these benefits may be problem specific and the increased computational overhead may not be justified in all situations.

Further Readings on Evolutionary Computing and the Genetic Algorithms

Cawsey A. *The essence of artificial intelligence.* London: Prentice-Hall, 1998.

Holland JH. *Adaptation in natural and artificial systems.* Ann Arbor, MI: The University of Michigan Press, 1975 (2nd ed '92).

Quagliarella D, Periaux J, Poloni C and Winter G. *Genetic algorithms and evolution strategies in engineering and computer science: Recent advances and industrial applications.* New York: John Wiley & Sons, 1998.

Darwin C. *On the origin of species by means of natural selection*. London: Murray, 1859.

Other Resources

Navy Center for Applied Research in Artificial Intelligence
http://www.aic.nrl.navy.mil/

The Genetic Algorithms Archive Page
http://www.aic.nrl.navy.mil/galist/

Appendix VI

Interaction terms included in LM1

SAS program used to generate interaction terms for LM1.

```
data ssd.dev;
set ssd.dev;
int1=albs*buns
;int2=albs*creats
;int3=albs*dbps
;int4=albs*fio2s
;int5=albs*glus
;int6=albs*hrs
;int7=albs*pao2s
;int8=albs*phs
;int9=albs*ptss
;int10=albs*rrs
;int11=albs*uos
;int12=albs*wbcs;
;int13=buns*creats
;int14=buns*dbps
;int15=buns*fio2s
;int16=buns*glus
;int17=buns*hrs
;int18=buns*pao2s
;int19=buns*phs
;int20=buns*ptss
;int21=buns*rrs
;int22=buns*uos
;int23=buns*wbcs;
;int24=creats*dbps
;int25=creats*fio2s
;int26=creats*glus
;int27=creats*hrs
;int28=creats*pao2s
;int29=creats*phs
;int30=creats*ptss
;int31=creats*rrs
;int32=creats*uos
;int33=creats*wbcs;
;int34=dbps*fio2s
;int35=dbps*glus
;int36=dbps*hrs
;int37=dbps*pao2s
;int38=dbps*phs
;int39=dbps*ptss
;int40=dbps*rrs
;int41=dbps*uos
;int42=dbps*wbcs;
;int43=fio2s*glus
;int44=fio2s*hrs
;int45=fio2s*pao2s
;int46=fio2s*phs
;int47=fio2s*ptss
;int48=fio2s*rrs
;int49=fio2s*uos
;int50=fio2s*wbcs;
;int51=glus*hrs
;int52=glus*pao2s
;int53=glus*phs
;int54=glus*ptss
;int55=glus*rrs
;int56=glus*uos
;int57=glus*wbcs;
;int58=hrs*pao2s
;int59=hrs*phs
;int60=hrs*ptss
;int61=hrs*rrs
;int62=hrs*uos
;int63=hrs*wbcs;
;int64=pao2s*phs
;int65=pao2s*ptss
;int66=pao2s*rrs
;int67=pao2s*uos
;int68=pao2s*wbcs;
;int69=phs*ptss
;int70=phs*rrs
;int71=phs*uos
;int72=phs*wbcs;
;int73=ptss*rrs
;int74=ptss*uos
;int75=ptss*wbcs;
;int76=rrs*uos
;int77=rrs*wbcs;
;int78=uos*wbcs;
output;
run
```

Appendix VII

Interaction terms included in LM2

```
data ssd.dev;
set ssd.dev;
int3s1=gcss3*albs3
;int3s2=gcss3*buns3
;int3s3=gcss3*creats3
;int3s4=gcss3*dbps3
;int3s5=gcss3*fio2s3
;int3s6=gcss3*hrs3
;int3s7=gcss3*nas3
;int3s8=gcss3*paco2s3
;int3s9=gcss3*pao2s3
;int3s10=gcss3*phs3
;int3s11=gcss3*ptss3
;int3s12=gcss3*rrs3
;int3s13=gcss3*temps3
;int3s14=gcss3*uos3
;int3s15=gcss3*wbcs3
;int3s16=albs3*buns3
;int3s17=albs3*creats3
;int3s18=albs3*dbps3
;int3s19=albs3*fio2s3
;int3s20=albs3*hrs3
;int3s21=albs3*nas3
;int3s22=albs3*paco2s3
;int3s23=albs3*pao2s3
;int3s24=albs3*phs3
;int3s25=albs3*ptss3
;int3s26=albs3*rrs3
;int3s27=albs3*temps3
;int3s28=albs3*uos3
;int3s29=albs3*wbcs3
;int3s30=buns3*creats3
;int3s31=buns3*dbps3
;int3s32=buns3*fio2s3
;int3s33=buns3*hrs3
;int3s34=buns3*nas3
;int3s35=buns3*paco2s3
;int3s36=buns3*pao2s3
;int3s37=buns3*phs3
;int3s38=buns3*ptss3
;int3s39=buns3*rrs3
;int3s40=buns3*temps3
;int3s41=buns3*uos3
;int3s42=buns3*wbcs3
;int3s43=creats3*dbps3
;int3s44=creats3*fio2s3
;int3s45=creats3*hrs3
;int3s46=creats3*nas3
;int3s47=creats3*paco2s3
;int3s48=creats3*pao2s3
;int3s49=creats3*phs3
;int3s50=creats3*ptss3
;int3s51=creats3*rrs3
;int3s52=creats3*temps3
;int3s53=creats3*uos3
;int3s54=creats3*wbcs3
;int3s55=dbps3*fio2s3
;int3s56=dbps3*hrs3
;int3s57=dbps3*nas3
;int3s58=dbps3*paco2s3
;int3s59=dbps3*pao2s3
;int3s60=dbps3*phs3
;int3s61=dbps3*ptss3
;int3s62=dbps3*rrs3
;int3s63=dbps3*temps3
;int3s64=dbps3*uos3
;int3s65=dbps3*wbcs3
;int3s66=fio2s3*hrs3
;int3s67=fio2s3*nas3
;int3s68=fio2s3*paco2s3
;int3s69=fio2s3*pao2s3
;int3s70=fio2s3*phs3
;int3s71=fio2s3*ptss3
;int3s72=fio2s3*rrs3
;int3s73=fio2s3*temps3
;int3s74=fio2s3*uos3
;int3s75=fio2s3*wbcs3
;int3s76=hrs3*nas3
;int3s77=hrs3*paco2s3
;int3s78=hrs3*pao2s3
;int3s79=hrs3*phs3
;int3s80=hrs3*ptss3
;int3s81=hrs3*rrs3
;int3s82=hrs3*temps3
;int3s83=hrs3*uos3
;int3s84=hrs3*wbcs3
;int3s85=paco2s3*pao2s3
;int3s86=paco2s3*phs3
;int3s87=paco2s3*ptss3
;int3s88=paco2s3*rrs3
;int3s89=paco2s3*temps3
;int3s90=paco2s3*uos3
;int3s91=paco2s3*wbcs3
;int3s92=pao2s3*phs3
;int3s93=pao2s3*ptss3
;int3s94=pao2s3*rrs3
;int3s95=pao2s3*temps3
;int3s96=pao2s3*uos3
;int3s97=pao2s3*wbcs3
;int3s98=phs3*ptss3
;int3s99=phs3*rrs3
;int3s100=phs3*temps3
;int3s101=phs3*uos3
;int3s102=phs3*wbcs3
;int3s103=ptss3*rrs3
;int3s104=ptss3*temps3
;int3s105=ptss3*uos
;int3s106=ptss3*wbcs3
;int3s107=temps3*rrs3
;int3s108=temps3*uos3
;int3s109=temps3*wbcs3
;int3s110=rrs3*uos3
;int3s111=rrs3*wbc3
;int3s112=uos3*wbcs3;
output;
run;
```

Appendix VIII

Pilot project 2: Comparing the ability of artificial neural networks and multivariate logistic regression to handle missing data.

Modeling mortality in the intensive care unit: comparing the ability of artificial neural networks and multivariate logistic regression to handle missing data.

Gordon S. Doig[1], Claudio M. Martin[2], Kevin J. Inman[2], James McD. Robertson[1] and William J. Sibbald[2]

[1]Department of Epidemiology and Biostatistics, University of Western Ontario,
[2]The Richard Ivey Critical Care Trauma Centre, Victoria Hospital Corp.,
London, Ontario, Canada

Running Title: Neural Nets and Missing Data

Contact person: Dr. Gordon S. Doig,
Department of Epidemiology and Biostatistics,
University of Western Ontario,
Ontario,
CANADA,
N6A 5C1
Phone: (519) 685-8500 ext 6308
E-mail: gdoig@biostats.uwo.ca

INTRODUCTION

In North America, the intensive care unit (ICU) accounts for seven percent of all hospital beds, fifteen to twenty percent of all hospital expenditures, and approximately one percent of the Gross National Product [Knaus *et al*., 1989]. Because the demand for intensive treatment is growing and resources are increasingly constrained [Schneiderman *et al*., 1990], it has become even more important to make effective decisions with respect to management practices and resource utilization. Recently much interest has been directed towards finding more effective tools to aid in the support of both clinical and management decisions in the ICU.

Alternative techniques such as nonlinear discrete neural networks have been successfully applied in solving some interesting medical problems. These techniques are derived from the engineering disciplines of pattern recognition and signals processing and are extremely promising because they offer the potential for ever-improving performance through dynamic learning [Levin, 1991].

A neural network trained on 351 patients with a high likelihood of myocardial infarction outperformed emergency room physicians when presented with 331 new cases of patients presenting with anterior chest pain. Attending clinicians diagnosed myocardial infarction with a sensitivity and specificity of 77% and 84% respectively, whereas the neural network performed with a sensitivity and specificity of 97.2% and 96.2% [Baxt, 1991].

Neural networks have outperformed clinicians on the diagnosis of hepatic masses [Maclin and Dempsey, 1992], pulmonary emboli [Scott and Palmer, 1993], and breast tumors [Goldberg *et al*., 1992]. Artificial neural networks have also shown their potential usefulness in the ICU by predicting the length of ICU stays after cardiac surgery [Tu and Guerriere, 1992].

Recently a neural network model was trained to predict mortality in the ICU. This model performed extremely well on its developmental dataset with an area under ROC curve of 0.9993. True predictive performance was compared with a multivariate logistic regression model created on the same developmental dataset by predicting the outcome of 138 admissions to which neither technique had previously been exposed. The true

predictive performance of each technique was found to be identical (area under ROC curve =0.82) [Doig *et al*., 1993].

The two reasons cited most often for neural networks' superior performance compared with more traditional techniques are their ability (i) to identify patterns of predictors not recognized by standard techniques and (ii) to predict accurately even with noisy or missing input data [Baxt, 1991; Tu and Guerriere, 1993]. Subsequent research has shown that neural networks do indeed have the ability to recognize patterns of predictors previously not associated with well investigated outcomes [Baxt, 1992], but none has investigated the ability of neural networks to predict accurately based on noisy or missing input values.

The purpose of this project was to compare and contrast the performance of a relatively simple back-propagation, associative-learning neural network with a classical multivariate logistic regression model based on the ability to predict mortality in an ICU given a validation dataset containing randomly generated noise in the form of missing values.

METHODS

Development of Predictive Models

The development and performance of the neural network and logistic regression models used in this simulation have been described in detail elsewhere [Doig *et al*., 1993]. They were developed on a prospectively collected dataset recorded on day 3 of admission to the intensive care unit to predict mortality during ICU stay.

Database Creation

During a six month period from August 5, 1991 to February 5, 1992, 614 patients were admitted to the 30 bed multi-disciplinary adult critical care unit. The only entry criterion for this study was a duration of stay greater than 72 hours. Four hundred and twenty-two patients met this criterion and were therefore entered into the study.

The 422 patients were randomly divided into a 2/3 developmental subset and a 1/3 validation subset. The 284 patient developmental dataset was used for the creation of both the neural network and the logistic regression models. The predictive performance of both models was then verified by predicting the outcomes of the patients in the 138 member

validation dataset. Neither logistic nor neural network models were exposed to any members of the validation datasets during model creation.

Input Variables

Fifteen variables were recorded daily for the duration of stay for each study entrant. These variables were identified from the literature [Knaus *et al*., 1985] and from clinical experience. They were: presence of acute renal failure, packed cell volume, heart rate, FIO_2, serum sodium, PaO_2, $PaCO_2$, pH, respiratory rate, systolic and diastolic blood pressures, serum potassium, temperature, white blood cell count, serum creatinine and the Glasgow coma scale score. The definition and recording of all variables was consistent with the methods outlined for data collection for the APACHE II scoring system [Knaus *et al*., 1985]. The outcome of interest was ICU mortality in patients with a duration of stay greater than 72 hours.

During the period of the study, the variables were abstracted from patient records and stored in a central codebook. At the conclusion of the study, the codebook was entered into a spreadsheet program and then transferred into PC SAS® version 6.04[1].

Database Validation

Primary data integrity was verified in PC SAS® with algorithms written to filter out biological impossibilities and obvious data entry transpositions. Any values in conflict with the screening filters were re-entered directly from the study codebook.

Secondary validation was carried out by match-merging the study database with a readily available XENIX based ICU management information system (MIS). This MIS database allowed validation of date of birth, ICU entry date, ICU discharge date and ICU discharge status. Since the MIS database is utilized for billing purposes, its entries are double-verified and seldom in error. Conflicts with the study database were resolved by accepting the MIS database as correct.

Primary validation revealed a coding error rate of 2.1% and independent secondary validation against the MIS database revealed an error rate of 2.4%. Comparison of study database errors against codebook values did not reveal any obvious transposition errors in the codebook. Only 0.2% of values were missing due to initial failure to perform laboratory tests at bedside.

Normal physiological values were imputed for the 0.2% of missing values, and both logistic and neural network models were created on identical datasets containing these imputed normal values.

Scale Selection

In order to improve convergence in the neural network and to reduce instability in the logistic model due to multicollinearity, scaling of the input variables was required. All variables except acute renal failure and the Glasgow coma scale score were transformed to the standardized normal distribution (z-scale).

Variables such as heart rate can convey different information about clinical interventions, outcomes and risks depending upon the degree of elevation or depression above or below the normal range. For this reason, all variables except serum creatinine, presence of acute renal failure and the Glasgow coma scale score were separated into high or low distributions about the mean. This resulted in a total of 27 variables, all of which were presented to the neural network and the logistic regression model building techniques.

Logistic Regression

Logistic regression was performed using PROC LOGIST, PC SAS®, version 6.04 [SAS Institute Inc., 1990]. Variables were considered as candidates for inclusion in the model based on a univariate logistic regression p-value 0.25 [Hosmer and Lemeshow, 1989]. A multiple-step backward model selection method was used and variables were removed from the model if significance fell above a p-value of 0.10. After the final model was evaluated, first order interaction terms were assessed. The final model contained seven significant main effects and no significant interactions (Table 1).

Neural Network

A commercially available back-propagation, associative-learning neural network was used for this simulation[2]. All 27 variables were presented to a 3 layered network with 27 input nodes, 18 hidden nodes and 1 output node. A logistic activation function was used and the output node generated a probability of mortality ranging between 0 and 1.

Through an error-minimization technique known as back-propagation, the neural network optimized weights between nodes such that important patterns between variables

were recognized. Using the 284 patient developmental database, the network converged on the optimum predictive solution after 20,300 iterations, which took 17:21 minutes on a 27 MHz 386.

A more detailed description of the development of the neural network and logistic regression models is provided elsewhere [Doig *et al*., 1993].

Generation of Missing Values

The complete database was composed of 422 patients × 15 input variables which resulted in 6330 elements. The validation dataset comprised 1/3 of the complete dataset (138 patients ×15 inputs). The 15 input variables were scaled and divided to create a total of 27 inputs to both the logistic regression and the neural network models.

Since validation of the original study dataset revealed an initial error rate of between 2.1-2.5%, it was decided that to simulate the worst case practice setting, this error rate should be inflated by a factor of two. Thus an absolute error rate of 5% missing values was generated.

To generate missing values in the validation dataset, a 27 column by 138 row binary transformation matrix was created. Each member of this matrix was independent and identically distributed, with a 95% probability of assuming a value of 1, and a 5% probability of assuming a value of 0. These probabilities were constrained by the Bernoulli distribution and generated using Quattro® Pro for Windows, Version 5.0[3].

Each element of the intact validation dataset was then multiplied by its corresponding element in the transformation matrix. When the transformation matrix element was equal to 1, the validation dataset element remained unchanged. When the transformation matrix element was equal to 0, multiplication resulted in a value of zero in the output matrix. A zero value in the input vector corresponds to a z-scaled 'normal' value. Thus missing values were represented by imputed normal values for the evaluation of both the neural network and the logistic regression model.

Statistical Comparisons

The neural network and the logistic regression model were compared based on their ability to predict mortality on the intact validation datasets. Next, their predictive

performance was compared based on their forecasts onto the validation dataset containing 5% corrupt values. The predictive performance of each technique was compared via the area under the ROC curve calculated for each technique [Erdich and Lee, 1981].

RESULTS

Patient Population

The average age of the study subjects was 61 years, the average duration of ICU stay was 7.3 days and the average day 1 APACHE II score was 25.3. The study population experienced a 9.5% mortality rate during ICU stay.

Intact Validation Dataset

At a classification threshold of 0.5, the logistic regression model performed with a sensitivity of 0.133 and a specificity of 0.976. The positive predictive value was 0.400 and the negative predictive value was 0.902. The area under the ROC curve was 0.8320.

At a classification threshold of 0.5, the neural network classified the 138 patient validation database with a sensitivity of 0.267 and a specificity of 0.976. The positive predictive value was 0.571 and the negative predictive value was 0.916. The area under the ROC curve was 0.8178.

Validation Dataset with 5% Missing Values

The logistic regression model performed with a sensitivity of 0.133 and a specificity of 0.967. The positive predictive value was 0.333 and the negative predictive value was 0.902. The area under the ROC curve was 0.804.

The neural network performed with a sensitivity of 0.200 and a specificity of 0.959. The positive predictive value was 0.375 and the negative predictive value was 0.908. The area under the ROC curve for the neural network was 0.800.

DISCUSSION

Although previous research has shown that neural networks are relatively insensitive to noise in the input parameters [Müller and Reinhardt, 1991], and that introducing noise into the input can actually improve performance in certain situations [Sondergaard, 1992], this simulation revealed that a relatively simple three-layered back-propagation network performed similarly to a multivariate logistic regression model (area under ROC curve = 0.80) when presented with a validation dataset with missing values

generated randomly at an overall rate of 5%. This is an extremely important finding, since the assumption that neural networks are insensitive to noise in the form of missing values could lead to a decreased vigilance during data collection, resulting in a loss in predictive performance.

Primary research into neural network architecture in the field of pattern recognition, where insensitivity to noise is extremely desirable, has yielded some interesting results. Lu and Szeto [1993] have shown that a network composed of subnets arranged in a hierarchical manner can perform extremely well in the presence of missing information. The subnets were trained to interpolate the appropriate value for the unknown information based on the known components of the input vector. This would be an extremely useful approach to dealing with missing information in biomedical systems, since it is known that certain clinical signs and symptoms do not present independently of each other.

For example, it has been widely accepted since the Roman era that the clinical signs of pain, swelling, heat and redness are the hallmarks of inflammation [Reference?, 450BC]. In any clinical case, swollen areas tend to be painful and reddened areas tend to be hot. Hence, the presentation of these four signs is not independent of each other. If the first three signs were present, that is, if a limb is *painful*, *swollen* and *hot* then imputing a 'normal' value of *not reddened* would probably be inappropriate. In this situation a hierarchical neural network, and also a clinician, would infer that the most appropriate value for the unknown sign, based on the three known signs, would be *reddened*. The presence of *reddening* would therefor have been predicated based on previous clinical experience and interpretation of the present clinical situation.

Researchers in fields other than clinical medicine have investigated application specific network architectures optimized to filter both random noise and gross errors in input [Rohwer *et al*., 1992; Kramer, 1992]. This pre-processed input data is then passed to the actual decision making engine. Simulations have show that this type of 'self-indexing' provides better results than assigning 'don't know' values to missing elements in the input dataset [Kak, 1993].

The type of neural network used in this simulation has previously been shown to compare favorably with a multivariate logistic regression approach to predicting mortality in the ICU. The three-layered back-propagation network has traditionally excelled at classification (pattern recognition) problems and is most useful in situations where the relationship between the input and the output is nonlinear and training data are abundant [Hinton, 1992]. The three or four-layered back-propagation network is also the type preferred by medical researchers. If the motivation for applying a similar network to a particular problem domain is the presence of noise in the form of missing input data, then more complex techniques should be considered.

References

Baxt WG. Analysis of the clinical variables driving decision in an artificial neural network trained to identify the presence of myocardial infarction. Ann Emerg Med 1992;21 :1439-1444.

Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med. 1991;115:843-848.

Chang RWS, Jacobs S and Lee B. Predicting outcome among intensive care unit patients using computerized trend analysis of daily Apache II scores corrected for organ system failure. Intensive Care Med 1988;14:558-566.

Civetta JM, Hudson-Civetta JA, Kirton O, Aragon C and Salas C. Further appraisal of APACHE II limitations and potential. Surg Gynecol Obstet 1992;175:195-203.

Doig GS, Inman KJ, Sibbald WJ, Martin CM and Robertson JMcD. Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. Proc Annu Symp Comput Appl Med Care 1993;361-365.

Erdreich LS and Lee ET. Use of relative operating characteristic analysis in epidemiology. Am J Epidemiol 1981;114:649-662.

Ferraris VA and Propp ME. Outcome in critical care patients: A multivariate study. Crit Care Med 1992;20:967-976.

Goldberg V, Manduca A, Ewert DL, Gisvold JJ, Greenleaf JF. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. Med Phys. 1992;19:1475-1481.

Hinton GE. How neural networks learn from experience. Sci Am 1992;267:145-151.

Hosmer DW and Lemeshow S. Applied Logistic Regression. John Wiley and Sons, New York, 1989.

Kak Sc. Feedback neural networks - new characteristics and a generalization. Circuits Systems And Signal Processing. 1993, V12, N2, P263-278

Knaus W, Wagner D and Draper E. APACHE III study design: analytic plan for evaluation of severity and outcome in intensive care unit patients. Implications. Crit Care Med. 1989a Dec. 17(12 Pt 2). P S219-21.

Knaus WA, Wagner DP, Draper EA, Simmerman JE, Bergner M, Bastos PG, Sirion CA, Murphy DJ, Lotring T, Damiano A and Harrel FE. The APACHE III Prognostic System; Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619-1636.

Knaus WA, Draper EA, Wagner DP, and Zimmerman. APACHE II: A severity of disease classification system. Crit Care Med 1985;13:818-830.

Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A, *et al.* The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991 Dec. 100(6). P 1619-36.

Kramer MA. Autoassociative neural networks. Computers & Chemical Engineering. 1992, Apr, V16, N4, P313-328.

Levin E, Gewirtzman R, Inbar GF. Neural network architecture for adaptive system modeling and control. Neural Networks 1991;4:2:185-191.

Lu SW and Szeto A. Hierarchical artificial neural networks for edge enhancement. Pattern Recognition. 1993, Aug, V26, N8, P1149-1163.

Maclin PS and Dempsey J. Using an artificial neural network to diagnose hepatic masses. J Med Syst 1992;16: 215-225.

Müller B and Reinhardt J. Physics of neural networks - an introduction. Springer-Verlag, New York, 1991.

Rohwer R, Grant B and Limb PR. Towards a connectionist reasoning system. Bt Technology Journal. 1992, Jul, V10, N3, P103-109.

Rutledge R, Fakhry SM, Rutherford EJ, Muakkassa F, Baker CC, Koruda M and Meyer AA. Acute Physiology and Chronic Health Evaluation (APACHE II) score and outcome in the surgical intensive care unit: an analysis of multiple intervention and outcome variables in 1,238 patients. Crit Care Med 1991;18:1048-1053

SAS® Institute Inc., SAS® User's Guide: Statistics, Version 6, 4th Edition, Cary, NC: SAS® Institute Inc. 1990.

Schneiderman LJ, Jecker NS and Jonsen AR. Medical futility: its meaning and ethical implications. Ann Intern Med 1990; 112:948-54.

Scott JA and Palmer EL. Neural network analysis of ventilation-perfusion lung scans. Radiology 1993;186:661-664.

Seneff M, Knaus WA. Predicting patient outcome from intensive care: a guide to APACHE, MPM, SAPS, PRISM and other prognostic scoring systems. J Intensive Care Med 1990;5:33-552

Sondergaard I, Krath BN and Hagerup M. Classification of crossed immunoelectrophoretic patterns using digital image processing and artificial neural networks. Electrophoresis. 1992 Jul. 13(7). P 411-5

Tu JV and Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Proc Annu Symp Comput Appl Med Care 1992; 666-672.

Tu JV and Guerriere MR. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Comp Biomed Res 1993;26:220-229.

Table 1	β	SE	OR	p
Hi Na	1.2876	0.34	3.624	0.000
Hi DBP	-1.8092	0.47	0.164	0.000
GCS	-0.2357	0.06	0.790	0.000
Hi PaCO2	-1.0222	0.50	0.360	0.042
Low Na	1.8762	0.64	6.529	0.003
Low K	2.0746	0.80	7.961	0.009
Low Temp	0.4771	0.22	1.611	0.029

β=regression parameter estimate, SE=standard error of regression parameter, OR=odds ratio, p= p-value

[1]PC SAS® version 6.04, SAS Institute Inc., SAS Circle, PO Box 8000,Cary, NC, 27512-8000, U.S.A.

[2]NeuroShell™,Ward Systems group, Inc., 245 W. Patrick St., Frederick, MD 21701, U.S.A.

[3]Borland International, Inc., 1800 Green Hills Road, P.O. box 660001, Scotts Valley, CA 95067-0001