Working Notes of the AIMDM'99 Workshop on

Prognostic Models in Medicine:
Artificial Intelligence and Decision Analytic Approaches

held during
The *Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*

Aalborg, Denmark, 20 June, 1999

Ameen Abu-Hanna & Peter Lucas (editors)

Preface

These are the working notes of the workshop on Prognostic Models in Medicine: Artificial Intelligence and Decision Analytic Approaches, which was held during the *Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99*, on 20 June, 1999, in Aalborg, Denmark. This workshop brought together various theoretical and practical approaches to prognosis that comprise the state of the art in this field. It is a follow-up on a very successful invited session on Intelligent Prognostic Methods in Medical Diagnosis and Treatment Planning during the conference *Computational Engineering in Systems Applications 1998* (CESA'98). AIMDM'99 joined the research fields of Medical Artificial Intelligence and Medical Decision Analysis. One of the aims of the present workshop was, therefore, to combine the views of researchers from these two different, although related, fields as well. This is not only reflected in the title of the workshop, but also in the contributions to it as contained in the working notes.

Prognostic models are increasingly used in medicine to predict the natural course of disease, or the expected outcome after treatment. In evaluating quality of care, prognostic models are used for predicting outcome, such as mortality, which may be compared with the actual measured outcome. Furthermore, prognostic models may play a role in guiding diagnostic problem solving, e.g. by only requesting information concerning tests, of which the outcome affects knowledge of the prognosis.

Various methods have been suggested in Medical Artificial Intelligence for the representations of prognostic models ranging from quantitative approaches, such as Bayesian networks and neural networks, to symbolic and qualitative ones, such as decision trees, as proposed within the machine-learning community. Dealing with semantic concepts such as time has been, and still is, a challenging issue. Temporal Bayesian networks and influence diagrams, and Markov decision processes have been developed as formalisms to deal with time explicitly. Similarly, in Medical Decision Analysis various representations with underlying techniques are suggested, such as decision trees, regression models, and representations in which advantage is taken from the Markov assumption. Hence, it is not easy to decide which representation formalism to choose to develop a specific prognostic model. The present working notes shed some light on this difficult issue, and offer a lot of useful practical experience in model building as well.

We are grateful to our colleagues who served on the programme committee of the workshop on Prognostic Models in Medicine (members are: A. Abu-Hanna (co-chair), S.S. Anand, S. Andreassen, P.M.M. Bossuyt, J. Fox, L.C. van der Gaag, J.D.F. Habbema, P. Haddawy, P. Hammond, E. Keravnou, N. Lavrač, J. van der Lei, P.J.F. Lucas (co-chair), L. Ohno-Machado, M. Ramoni, M. Stefanelli, Th. Wetter, J. Wyatt). They have carefully read and reviewed each submission until the acceptance of the final papers. Thanks are also due to Dik Habbema, Kristian Olesen and Jeremy Wyatt for accepting to give the three invited talks of the workshop.

Ameen Abu-Hanna, Department of Medical Informatics, University of Amsterdam
Peter Lucas, Department of Computer Science, Utrecht University
6 June, 1999

Contents

Invited Talks

Building Prognostic Models: statistical aspects

Dik Habbema & Ewout Steyerberg
Department of Public Health
Medical Faculty, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
Email: habbema@mgz.fgg.eur.nl

Prediction is a crucial activity in clinical medicine. Sometimes it deals with the unknown future; in this case we want to predict the outcome for patients ('prognosis'). Alternatively, prediction concerns the present, namely prediction of disease condition or diagnosis (the 'pre-' in prediction refers here to 'before you know'). We will use the terms predictors and outcome for the statistical variables involved; alternative names are explanatory variables and explicandum, or independent variables and dependent variable etc.

Two issues will be addressed:

- Which predictors to use?
- How to measure the performance of the prediction rule?

Which predictors to use? We will question the general usefulness of trying to select a few predictors out of a larger number available predictors. Recent results indicate that classical statistical forward or backward selection procedures have been used with too much enthusiasm. Problems will be discussed, and illustrated with examples. Sample size is important, and cautiousness is especially indicated for small samples. A closely connected question is what importance to attach to the selected versus the non-selected predictors. Here a lot of confusion exists, and we as methodologists may have been too lazy with explaining to clinicians that there may be a high degree of arbitrariness in the result of the selection process, and that further inspection is needed before equating 'selection' to 'importance'.

This will again be discussed using examples, and some emphasized on analyzing the structure of the mutual dependency between the predictors, and the dependency between predictors and predictor.

Performance measurement, which is also relevant for the selection of predictors issue, will be taken further. A well known fundamental observation is that a too favorable impression of performance is obtained when it is measured on the same patients as have been used for the construction of the prediction model. There are two main approaches to more realistic estimation of regression coefficients and performance measurement: a simulation approach using bootstrap re-sampling and a formula-based direct shrinkage of regression coefficients. In large samples, split sample may also be used: although this method is easily understandable to clinicians,we should realize that information will be thrown away.

Finally, some remarks will be made on differences between exploratory, explanatory and predictive use of the type of models as discussed. It is argued that the basic process will

remain the same under these three uses, but emphasis and detail may differ considerably. When clinical use of the prediction model is aimed at, the costs of measuring the predictors have to be taken into account, at least when they have no other clinical relevance than their use in the prediction model.

Medical Models and Bayesian Networks

Kristian G. Olesen
Department of Computer Science, Aalborg University
Frederik Bajers Vej 7, DK-9220 Aalborg Ø, Denmark
Email: kgo@cs.auc.dk

The use of computers in health care has increased dramatically over the past decades and computers are now involved in practically all tasks in the sector, including patient management, communication, diagnosing, planning of theraphy and test, prognosis, and quality control. The first two tasks mentioned in the list are well understood and reliable computer support for them exists. The remaining tasks are more complicated and, in general, characterized by inherent uncertainty, and this makes it difficult to construct generally accepted automated tools to support them. It is important that such systems are transparent and that they meet their users on the users premises. Models based on the concepts usually used by physisians are preferred and such models should exhibit the structure of the domain in question through explicit representation of the relations between concepts. This enables a scientific discussion of the model itself independent of its use, thereby making it transparent and open for criticism. Moreover, such models should be powerful enough to support procedures for a variety of tasks. Bayesian networks offer a framework in which such models can be constructed.

This talk will present the basics of Bayesian networks and illustrate their use through a number of examples. The basic structure of a Bayesian network model consists of a graph where the nodes are stochastic variables modelling the concepts of a domain. The edges of the graph models direct dependencies between variables and the strength of these dependencies are quantified by conditional probability distributions. Bayesian networks represent the joint probability distribution over all variables in the domain and this distribution can be updated dynamically as evidence arrives. This makes Bayesian networks suited for diagnostic systems where the current belief in a set of disorders can be maintained, thereby representing the basis for decisions.

The planning of tests and therapy involve active decisions, where the state of the world is influenced. These tasks can be integrated in the Bayesian network formalism through the addition of decision and utility nodes. Such extended models are known as influence diagrams. Decision nodes are under the full control of the decision maker and utility nodes are real-valued functions used to model the decision makers preferences. Optimal decisions are then computed by maximizing the expected utility.

Prognostic Models in Medicine

J.C. Wyatt
School of Public Policy
University College London
29/30 Tavistock Square
London, WC1H 9EZ, UK

Papers

Prognostic Models in Medicine: Artificial Intelligence and Decision Analytic Approaches

Peter Lucas
Department of Computer Science
Utrecht University, PO Box 80089
3508 TB Utrecht, The Netherlands
E-mail: lucas@cs.uu.nl

Ameen Abu-Hanna
Department of Medical Informatics
AMC-UvA, Meibergdreef 15
1105 AZ Amsterdam, The Netherlands
E-mail: a.abu-hanna@amc.uva.nl

Abstract

This paper is meant as an introduction to the workshop on *Prognostic Models in Medicine: Artificial Intelligence and Decision Analytic Approaches* held during AIMDM'99. Prognosis – the prediction of the course and outcome of disease processes, either or not changed due to interventions – is an important aspect of medical tasks like diagnosis and treatment management. Techniques for building prognostic models vary from traditional probabilistic approaches, originating from the field of statistics, as used in decision analysis, to more qualitative and model-based approaches originating from the field of artificial intelligence. The workshop brings these two fields of research together in the hope that a fruitful exchange in ideas will take place.

1 Introduction

Prognosis, the prediction of the course and outcome of disease, is a subject that lies at the heart of patient management. There is little sense in delving into the cause of particular symptoms and signs in a patient, and to initiate elaborate diagnostic procedures, if it is known beforehand that no effective treatment of the considered disease exists. Furthermore, also treatment selection invariably involves taking possible future beneficial and harmful effects into account, i.e. prognostic information [4].

Of course, the process of patient management concerns issues other than prognosis as well. The primary role of the physician is to guide the patient through the disease process, which involves much more than prognostication. Even the processes of diagnosis and treatment selection may be seen in this light of guidance of patients. This view may explain why prognosis, despite its central role in medicine, is not clearly recognised as such in typical medical textbooks, like *Harrison's Principles of Internal Medicine* [1]. The subject of prognosis is only paid attention to when it is obviously important, such as in cancer treatment.

It is likely that this situation will change in the near future, and that the role of prognostic models in medicine will increase. Medicine as a field is becoming increasingly complex, as is reflected by the annually increasing number of different diagnostic tests and therapies from which a clinician must choose. Prognostic models are required to guide clinicians in this selection process to ensure that the patient will benefit from further progress in medical science.

2 Prognostic models

As has been said above, there are a number of fields in medicine, where prognostic models are of particular importance. Examples of such fields are: oncology, transplantation medicine, and trauma medicine. Usually, prognostic models focus either on long-term or short-term effects. For example, long-term effects dominate in treatment considerations in oncology, whereas short-term effects are more significant in trauma medicine. Adequate prognostic information is of major importance in these fields so that prognostic models of various kinds, not necessarily mathematical in nature, have been in use for quite some time. Often these models are coarse and lack detail. The TNM staging system that is used to assess a primary malignant tumour in terms of its size (indicated by T_0 to T_4, where an increase in subscript corresponds to an increase in tumour size, as defined for a particular type of tumour), regional lymph node involvement (N, also supplied with a subscript), and presence of distant metastatis (M) is an example of a simple qualitative tool to assess prognosis in cancer patients. Another example is the Apache III scoring system, which is based on a logistic regression model, and that has been shown to have a good predictive ability for patients with severe illness, and for a large variety of diseases [2].

As one may expect, clinicians are only prepared to accept prognostic models when it is obviously that they will contribute to quality of care [7]. Prognostic models are not only used in a clinical setting. They are also used, and may even have had a larger impact, in the design of clinical trials, counselling patients and in medical technology assessment.

In general, and independent of particular applications of prognostic models, the problem of the design of accurate prognostic models is the capturing of the many possible subtle interactions among variables that exist. It is largely determined by the (mathematical) modelling tools used to what extent such interactions can be represented, and learnt from data, possibly augmented with background knowledge.

3 Artificial intelligence and decision analysis

Medical artificial intelligence is generally concerned with the development of medical models for various purposes, but usually the aim is to assist clinicians in the processes of diagnosis, treatment or prognosis of diseases in patients. A key characteristic is the *explicit* representation of the medical knowledge involved, i.e. the explicit representation of meaningful interactions among the factors that play a role in a particular medical problem is favoured [3]. However, there are a number fields in artificial intelligence, such as neural networks, where the goal of explicit representation is less dominant. There is now an entire array of different techniques from which medical AI practitioners may choose. One of the difficult problems has been the representation of temporal patterns, which is now addressed by a number of different formalisms. Progress in the field has yielded new, flexible techniques, like Bayesian networks, neural networks and genetic algorithms; these offer new opportunities for dealing with the issue of prognostication.

Medical decision analysis offers a systematic approach to medical decision making under conditions of uncertainty [5; 6]. It has studied the use of prognostic models in the process of decision making for more than two decades. An enormous amount of practical experience in building medical models has been built up during these years. However, there has been little progress in the field with respect to new techniques and tools that may be used to carry out a decision analysis.

Until recently the fields of medical artificial intelligence and decision analysis appeared to have only in common that in both model building is of crucial importance. In the field of artificial intelligence there has been a revival of interest in numerical methods stemming from probability and decision theory, and from the field of neural networks. The new ideas and techniques that have come out of this, has not passed by unnoticed by the medical decision analysis community. There currently seem to be much interest in that field with respect to applicability of these technique. At the same time, medical artificial-intelligence researchers realise that much can be learnt from the more mature field of medical decision analysis. This workshop is therefore a timely opportunity to exchange ideas and hopefully to learn from each other.

4 Road-map to the workshop papers

To conclude this introductory paper, we shall briefly summarise the contents of the papers in the working notes.

The paper by S.S. Anand, P.W. Hamilton, J.G. Hughes and D.A. Bell, titled *Utilising censored neighbours in prognostication*, discusses an extended version of the k-nearest neighbour algorithm, which is applied to the problem of prediction of the survival of patients with colorectal cancer. Novel is the possibility of dealing with censored patients, which is typically required in survival analysis in medicine. The paper by S. Antel, L.M. Li, F. Cendes, Z. Caramanos, A. Olivier, F. Andermann, F. Dubeau, R.E. Kearney, R. Shinghai and D.L. Arnold, with title *A naive Bayesian classifier for the prediction of surgical outcome in patients with temporal lobe epilepsy*, focusses on a number of important issues that arise when one wants to develop prognostic models the are clinically useful. The development of a Bayesian classifier for the prediction of the outcome of patients with temporal lobe epilepsy that undergo surgery is reported. Bayesian classifiers are also the topic of the paper *Robust outcome prediction for intensive-care patients* by M. Ramoni, P. Sebastian and R. Dybowski, but here the main issue is how to deal with missing values in clinical data. A comparison is made between logistic regresssion augmented with an imputation mechanism and what is called a *robust* Bayesian classifier in which no assumptions are made with respect to the mechanisms underlying missing data.

In the paper by H. Dreau, I. Colombet, P. Degoulet, G. Chatellieri, titled *Identification of patients at high cardiovascular risk using a critical appraisal of statistical risk prediction models* not techniques, but different statistical risk-prediction models are compared. This paper sheds light on the assumptions underlying statistical models, and on the question to which extent assumptions are valid and may affect the conclusions that may be drawn.

There are a number of papers in which statistical or decision-analytic techniques are compared or combined with AI techniques. For example in the paper by L. Ohno-Machado and S. Vinterbo, *Influential case detection in medical prognosis* it is studied whether a genetic algorithm offers advantages over conventional techniques for the selection of cases in the construction of prognostic logistic regression models. The prediction of the prognosis of trauma patients has been chosen as an example domain. I. Zelič, N. Lavrač, P. Najdenov and Z. Rener-Prime in their paper *Impact of machine learning to the diagnosis and prognosis of first cerebral paroxysm* compare ID3-like decision-tree induction with naive Bayesian classifiers from the perspective of machine learning. The comparison is carried out in the medical domain of epilepsy.

The remaining two papers focus on medical applications of techniques from the areas of artificial intelligence. In the paper by N. Peek, *A specialised POMDP*

form and algorithm for clinical patient management the formalism of partially observable Markov decision problems (POMDPs) is studied. This formalism has originally been introduced in artificial intelligence as a means to handle planning problems under conditions of uncertainty. POMDPs, however, have also been suggested as a suitable formalism for medical treatment planning. Since the formalism is known to be intractable in general, this papers proposes the use of Monte Carlo simulation to render the formalism practically more useful. The paper by R. Schmidt, B. Pollwein and L. Gierl, titled *Prognoses for multiparametric time course of the kidney function* also discusses the suitability of a technique from the field of artificial intelligence to the development of prognostic models, namely the application of case-based reasoning to the prediction of kidney function. Advantages and limitations of case-based reasoning are clearly discussed.

We may conclude that the papers in the workshop, although all dealing with the issue of prognostic mocels, are indeed varied; both methods and techniques from the fields of artificial intelligence, decision analysis and statistics are covered by the papers. Sometimes these techniques are dealt with separately, sometimes they are combined and in some papers they are compared to each other. It may therefore be concluded that the title of the workshop does indeed reflect the contents of the papers in the workshop notes.

References

[1] K.J. Isselbacher, et al., *Harrison's Principles of Internal Medicine*, 13th edition, McGraw-Hill, New-York, 1994.

[2] W.A. Knaus, E.A. Draper, J. Lynn, Short-term morbidity predictions for critically ill hospitalised patients – science and ethics, *Science* 254 (1991) 389–94.

[3] P.J.F. Lucas, Logic engineering in medicine, The Knowledge Engineering Review, Vol. 10, No. 2, 1995, pp. 153–179.

[4] P.J.F. Lucas and A. Abu-Hanna, Prognostic methods in medicine. *Artificial Intelligence in Medicine*, vol. 15, 1999, pp. 105–119.

[5] H.C. Sox, M.A. Blatt, M.C. Higgins, K.I. Marton, *Medical Decision Making*, Butterworths, Boston, 1988.

[6] M.C. Weinstein and H.V. Fineberg, *Clinical Decision Analysis*, W.B. Saunders, Philadelphia, 1980.

[7] J.C. Wyatt and D.G. Altman, Commentary – prognostic models: clinically useful or quickly forgotten, BMJ, Vol. 311, 1995, pp. 1539–1541.

Utilising Censored Neighbours in Prognostication

Sarabjot S. Anand, John G. Hughes, David A. Bell
School of Information and Software Engineering,
University of Ulster at Jordanstown,
Newtownabbey, County Antrim
Northern Ireland BT37 0QB

Peter W. Hamilton
Department of Pathology,
Queens University of Belfast,
Northern Ireland

Abstract

Evaluation of new modelling techniques is an essential part of their development and acceptance within the medical domain. The models developed need to be evaluated along a number of dimensions - accuracy of the resulting model, perspicuity of the model, its ability to handle domain knowledge, its ability to handle data specific characteristics and its ability to continually refine the model. Feedback from such evaluations must be used to enhance present modelling techniques. In this paper we extend the basic k-nearest neighbour (k-NN) paradigm, based on an earlier evaluation [Anand et al., 1999], so as to enhance its capabilities with respect to handling censored patient data. We refer to this new k-NN algorithm as *Censored k-NN* (Ck-NN). Two aspects of the k-NN are extended, the distance metric, used to retrieve the nearest neighbours of a target case, and the prediction mechanism used to provide a point estimate for the dependent attribute. Ck-NN is evaluated by using it to model survival time for colorectal cancer patients.

1 Introduction

Prognostic models have traditionally been developed using methods from medical statistics that have been developed to handle complexities within medical data. Censored observations are only one such aspect of medical data that make modelling more complex. Statistical techniques such as Cox's regression, Kaplan-Meier and Weibull modelling [Collett, 1994] deal with such data.

It is our belief that new techniques used to build prognostic models, just as in the case of any other form of decision support system in medicine, must provide additionality - net added benefits - over these traditional, established techniques if they are to find large-scale acceptance within the medical domain. In general, additionality must be measured along a number of different dimensions – accuracy of the resulting model, perspicuity of the model, its ability to handle domain knowledge, its ability to handle data-specific characteristics such as skewness of distributions, censored observations etc. and its ability to continually refine the model, providing support for both the generic learning tasks of knowledge acquisition and knowledge refinement.

In a previous paper, we compared neural networks, regression tree induction, linear regression, Cox's regression and the nearest neighbour paradigm with the aim of ascertaining their suitability for modelling survival time for colorectal cancer patients [Anand et al., 1999]. The conclusions arrived at in that study can be summarised as follows. Firstly, there is a general lack, in the literature, of evaluation of new AI techniques against tried and tested statistical techniques. Secondly, within the domain addressed, Cox's regression and neural networks achieved similar accuracy. However, neural networks were, in general, unable to handle censored patient data. The work by Farragi and Simon [Faraggi and Simon, 1995] is an exception to this rule. With respect to perspicuity of the model, both neural networks and Cox's regression are not ideal, as the survival baseline and exponential terms in Cox's model reduce the intuitiveness of the interpretation of the model while neural networks are known to generate a "nervousness" within clinicians [Wyatt, 1995]. In fact, it has been pointed out that regression trees that are generally thought of as being readily understandable fail to meet another aspect of perspicuity - intuitiveness [Lavrač, 1998]. Clinicians find regression trees to be less intuitive as they utilise minimal "relevant" information as opposed to the information available in its entirety. The nearest neighbour paradigm was found to be psychologically plausible and understandable. However, the basic paradigm scored low on the accuracy scale mainly due to the presence of irrelevant attributes and the existence of biases within the distance metric used. Another failing was its inability to handle censored observations. Enhancements made to the basic paradigm with respect to the distance metric and attribute weight discovery mechanism, proved to be effective in improving the accuracy of the model. The issue of handling censored data within the nearest neighbour paradigm was only briefly discussed in previous work by the authors, and is the subject of the present study. More

specifically, we investigate how censored observations can be incorporated into the distance metric and how the censored and uncensored retrieved nearest neighbours can be utilised to arrive at a single prediction for the dependent attribute of the target example.

2 Incorporating Censored Data within the Distance Metric

Traditionally, the k-NN uses the Euclidean and Manhattan distances to compute the distance between the target example and exemplars within the exemplar base. When some of the attributes describing the exemplars are categorical, these traditional metrics introduce a bias, into the retrieval of the nearest neighbours, towards matching categorical attribute values. Anand et al. [Anand and Hughes, 1998] introduced a number of enhanced distance metrics that remove this bias when using the nearest neighbour paradigm for predicting a continuous valued dependent attribute, such as, survival time.

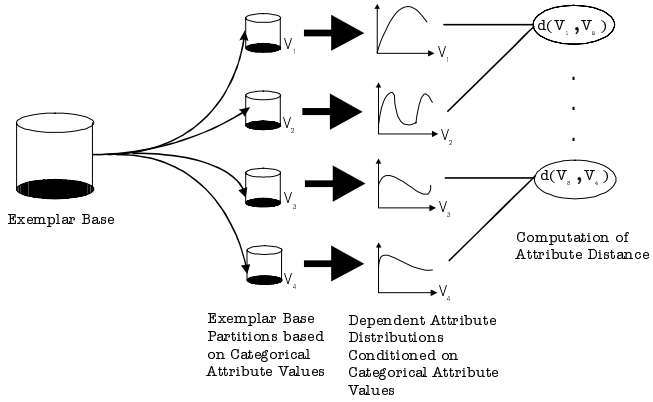

Figure 1: Enhanced Distance Metrics for Categorical Attributes

The enhanced distance metrics are based on the following observation (Figure 1). A categorical attribute, with domain say $\{V_1, V_2, V_3, V_4\}$, effectively partitions the exemplar base. Each partition is defined by a unique categorical attribute value. We refer to the distribution of the dependent attribute within a partition as the distribution conditioned on the categorical value defining the partition. Anand et al. define the distance between two categorical attribute values, say $d(V_1,V_2)$, based on the difference between the resulting dependent attribute distributions conditioned on the values of the categorical attribute. The difference between two distributions may be defined by using any of a number of statistical tests, comparing either the central tendency measures of the two distributions (for example, using the t-test), or the whole distributions (for example, using the Kolmogorov-Smirnov test).

While such a definition of the distance between categorical values has been shown to be effective [Anand et al., 1999], in the presence of censored observations, the observed survival time distribution may be quite different from the true distribution and the resulting distance metric would once again be biased. In this section we discuss an alternative definition of the distance between two categorical values, based on the survivor curve, as defined in the Kaplan-Meier model for survival analysis, conditioned on the categorical values. The advantage of using the survivor curves rather than the survival time distributions is that the survivor curves take account of censored observations in their definition, whereas the survival time distributions do not do so.

The enhanced metrics defined in [Anand and Hughes, 1998] as well as the metrics described in this section can be justified by the fact that the nearest neighbour assumes independence of the attributes used to build the model (as in the case of the naive Bayes methods). Based on the independence assumption the distribution of survival times conditioned on a categorical value can be assumed to be unaffected by any of the other attributes describing the exemplar. Thus, the Kaplan-Meier based survivor curve definition is sufficient. For cases were this does not hold, Kasif et al. suggest a probabilistic framework for the nearest neighbour paradigm that may be employed [Kasif et al., 1998].

The survivor curves corresponding to two different values of a categorical attribute (conditional survivor curves) may be compared in a number of ways. A simplistic method is to compare the survival time for the two curves at the half-life (i.e. survivor probability of 0.5). Alternatively, more rigorous methods of comparison such as the log-rank test and Wilcoxon test may be used. We investigate all three methods here.

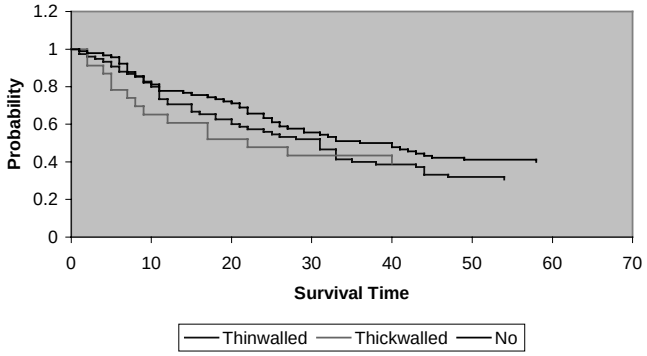

Figure 2: Survivor curves conditioned on values of Venous Invasion

Figure 2 shows the conditional survivor curves for Venous Invasion using a data set from the colorectal cancer domain [Anand et al., 1999]. The resulting mapping of the categorical attribute values onto a numeric scale of [0,1] using the median survival time (i.e. the survival time at probability 0.5), are shown in Table 1. Table 1 also shows the mapping of Venous Invasion when using the Mean and Coefficient of Variation as the basis for the mapping discussed previously [Anand and Hughes, 1998]. The mapping onto the [0,1] scale based on the median survival is medically more intuitive than those based on central tendency measures of the survival distribution itself. As would be expected, the distance between no venous invasion and

thick-walled venous invasion is greater than the distance between thin-walled venous invasion and the two extremes.

Venous Invasion	*Median Survival*	*Mapped Value*	*Mean Based*	*Coefficient of Variation based*
No	36	1	1	0.223
Thin-walled	31	0.687	0	0
Thick-walled	22	0	0.172	1

Table 1: Mapping of Venous Invasion onto the scale [0,1]

The Log-Rank test and Wilcoxon test are two non-parametric tests that can be used as a more rigorous basis for the comparison of survivor curves. Both statistics are distributed as the chi-square distribution with one degree of freedom. Thus, the probability, p, that the two curves belong to the same underlying distribution can be obtained using the chi-square distribution. Using the Log-Rank test and Wilcoxon test, the resulting distance matrices, defined as 1-p, are as shown below.

	Thinwalled	
No	0.211	No
Thickwalled	0.818	0.703

Table 2: Distance Matrix for Venous Invasion using the Log-Rank test

	Thinwalled	
No	0.533	No
Thickwalled	0.807	0.855

Table 3: Distance Matrix for Venous Invasion using the Wilcoxon test

In the case of the Log-Rank test the resulting distances do not follow the intuitive expectations as in the case of the median based and Wilcoxon test based metrics. Further investigation of why this is the case must be undertaken. One possible explanation is that the Log-rank test is more sensitive to changes in the tail of the left-skewed curves which may be causing this anomaly.

3 Predictions based on Censored Patient Data

Using the distance metrics defined in the previous section, the 'k' nearest neighbours for a given target are retrieved from the exemplar base. These exemplars must now be used to arrive at a single prediction for the target.

In the case of the colorectal cancer data set, all censored observations are right censored. Thus, the recorded observed survival for the censored observations can be interpreted as lower bounds to the true survival lengths for these patients and may be modelled within the data as a survival of "greater than <observed value>" (Table 4). In effect, what we now have is a complex, structured dependent attribute that needs to be modelled using the existing set of independent attributes. While such an approach is intuitive, few modelling techniques can handle such a structured dependent attribute.

Consider the case where k nearest neighbours $n_1, n_2, \ldots, n_k$ have been retrieved from the exemplar base when presented with the target example, t. Let $o_1, o_2, \ldots, o_k$ be the dependent attribute values for the k neighbours and $d_1, d_2, \ldots, d_k$ be the distances of these k neighbours from the target. Of the k neighbours, let us assume without loss of generality, that the first c neighbours are uncensored observations while the rest of the k-c neighbours are censored. Now, using the kernel function, K(d), defined below, we can associate a vote with each of the retrieved neighbours denoted by $v_1, v_2, \ldots, v_k$. The kernel function should associate a smaller vote with neighbours that are further away from the target.

$$v_i = K(d_i) = \frac{1 - d_i / \sum_k d_j}{\sum_k \left(1 - d_i / \sum_k d_j\right)}$$

Sex	*Path Type*	*Polarity*	*Configuration*	*Pattern*	*Infiltration*	*Fibrosis*	*Venous Invasion*	*Mitotic Count*	*Penetration*	*Differentiation*	*Dukes Stage*	*Age*	*Obstruction*	*Site*	*Survival*
M	Signet	lost	simple	expanding	little	little	no	0-5	node	well	A1	34	No	rectum	> 40
F	Tubular	Easily discerned	simple	infiltrating	marked	little	no	6-10	serosa	poor	C	48	No	caecum	> 22
M	Signet	Easily discerned	simple	expanding	little	little	yes	0-5	node	poor	D	78	No	caecum	5

Table 4: Example data using "lower bound" interpretation of right censored data

The votes $v_1, v_2, \ldots, v_c$ can be used to combine the values of the dependent attribute associated with the c uncensored observations using the formula:

$$o = \frac{\sum_{i=1}^{c} v_i . o_i}{\sum_{i=1}^{c} v_i}$$

with an associated vote v defined as:

$$v = \sum_{i=1}^{c} v_i$$

.

Now, given the definition of the kernel function, it is straightforward to prove that the kernel function is an evidential mass function [Anand et al., 1996]. The function, m, that maps the k-c+1 dependent attribute values onto the interval [0,1] representing the evidential mass associated with that particular value of the dependent attribute being the expected value may be defined as:

$m(o_j) = v_j \ \forall\ i \in [k\text{-}c, k]$
and $m(o) = v$

Associated with the mass function, a belief function may be defined as,

$$bel(X) = \sum_{Y \subseteq X} m(Y)$$

where X and Y are subsets of the frame of discernment defined as the set of all possible outcome values.

The outcome value with belief greater than and closest to 0.5 would be the preferred outcome. The value of 0.5 is the generally accepted value used in modelling. In survival analysis it is often referred to as the half-life, as it is the point at which the probability of the predicted outcome being the true outcome is not less than the probability of the predicted outcome being incorrect. A more conservative prediction would be to use a higher threshold value. The value of 0.5 is taken to be a balance between correctness and informativeness of the predicted value. For example, a prediction of "> 0" will always be a correct prediction but its informativeness will be the lowest possible.

Neighbour #	*distance from target*	*dependent attribute value*	*Vote (using kernel function)*
1	1.4	25	0.207
2	1.5	30	0.204
3	1.6	> 30	0.201
4	1.8	> 40	0.195
5	2	> 50	0.189

Table 5: Example retrieved neighbours

We illustrate the prediction mechanism using an example, assuming k to be 5. Table 5 shows the distance for the target, dependent attribute values and votes associated with each of the five neighbours. In the example, there are three censored observations within the retrieved set of neighbours. Combining the uncensored observations we get a combined value of 27.481 and vote of 0.411. The resulting mass function is:
m(27.481) = 0.411, m(> 30) = 0.201, m(> 40) = 0.195, m(> 50) =0.189.
and the associated belief function is:
bel(27.481) = 0.411, bel(> 30) = 0.585, bel(> 40) = 0.384, m(> 50) = 0.189.

Seeing that the belief associated with the dependent attribute value of '> 30' is greater than and closest to 0.5, using half-life as the threshold, we can predict for the target example that the survival value will be greater than 30 months.

4 Evaluation

Evaluation of predictive models in the presence of censored observation poses a number of problems. Unlike, cases where no censored observations exist, simply using the mean absolute error in prediction as a measure of accuracy is not good enough [Anand et al., 1999]. In fact, the mean absolute error measures the informativeness but not the accuracy in the presence of censored observations. In this section we report on some preliminary results obtained using various enhanced distance metrics described previously [Anand et al., 1999] and Ck-NN.

Distance Metric	*Type 1*	*Type 2*	*Type 3*	*Type 4*
Euclidean	19.42 (105)	45.19 (0/46)	43.86 (15/15)	26.35 (6/22)
Significant Mean	19.42 (105)	45.19 (0/46)	43.86 (15/15)	25.95 (4/22)
Mean	21.99 (105)	42.97 (0/44)	46.73 (15/15)	29.62 (8/22)
Coefficient of Dispersion	17.99 (107)	44.94 (0/50)	48.61 (13/13)	29.66 (3/18)
Censored Median	10.03 (98)	62 (0/45)	50.09 (22/22)	22.26 (7/23)
Log Rank	10.32 (100)	62.93 (0/44)	43.95 (20/20)	25.45 (4/24)
Wilcoxon	10.32 (100)	62.93 (0/44)	43.95 (20/20)	25.41 (5/24)

Table 6: Results in the absence on attribute weights

We define four types of prediction outcomes. Type 1 is where both the predicted and actual values are uncensored, Type 2 is where the predicted value is uncensored and actual value is censored, Type 3 is where the predicted value is censored and the actual value is uncensored and, finally, Type 4, where both values are censored. Table 6 shows the mean absolute error for each type of outcome, using 10-fold cross validation and the number of observations for which the predicted value is greater than the actual value from Type 2, 3 and 4 (in brackets). Clearly, the higher the number in the brackets the better for Type 2 outcomes and a lower value for Type 3 and 4 are preferred. In all cases a lower mean absolute value is clearly desirable. As would be expected, the MAE in Type 1 predictions has decreased sub-

stantially when using the Ck-NN method. For Type 2 predictions the MAE is least informative as the observed value is censored and the observed value may be closer to the predicted value. Type 3 and 4 seem unaffected by the new method.

Distance Metric	*Type 1*	*Type 2*	*Type 3*	*Type 4*
Euclidean	15.79 (112)	40.71 (0/45)	43.87 (8/8)	26.56 (2/23)
Significant Mean	16.25 (111)	42.55 (0/47)	45.55 (9/9)	24.33 (3/21)
Mean	19.875 (104)	42.09 (0/44)	39 (16/16)	25.39 (7/24)
Coefficient of Dispersion	17.03 (104)	42.15 (0/48)	39.75 (16/16)	27.5 (2/20)
Censored Median	10.01 (92)	57.95 (0/44)	43.14 (28/28)	29.08 (5/24)
Log Rank	9.98 (108)	57.94 (0/37)	44 (12/12)	26.06 (6/31)
Wilcoxon	9.98 (108)	57.94 (0/37)	44 (12/12)	26.06 (8/31)

Table 7: Results in the absence on attribute weights

A well known drawback of the k-NN approach is its sensitivity to irrelevant attributes, the data used in the evaluation consists of 188 examples representing patients diagnosed with colorectal cancer. Fifteen clinico-pathological attributes were recorded for each patient. Previous studies have shown that the relevance of these attributes vary widely and genetic algorithms have been employed for discovering optimal attribute weights [Anand and Hughes, 1998]. In the absence of censored observations the genetic algorithm uses as its fitness function, the mean absolute error obtained when using the k-NN algorithm with the feature weights represented by a "chromosome" in the genetic pool. In this paper we use the same fitness measure as minimising the mean absolute error optimises the informativenss of the predictions. Table 7 shows the results obtained. As can be seen from Table 6 and 7, using attribute weights decreases the mean absolute error of the model. Also, the number of examples of Type 1 and 4 have increased while the numbers in Type 2 and 3 have reduced.

5 Concluding Remarks

The inability to handle censored observations within artificial intelligence systems is just one of the obstacles in their path towards acceptance as methods for enabling medical decision support. In this paper, we outlined a methodology (Ck-NN) that may be employed to incorporate such observations within the similarity or distance metric and the method for arriving at a single prediction within the nearest neighbour paradigm. Ck-NN utilises the Kaplan-Meier estimate for the survivor curve to derive a distance/similarity metric for categorical attributes. It then employs elements of evidence theory to represent the evidence gathered, from the retrieved nearest neighbours, about the probable survival time of patients.

A rigorous evaluation of the methodology using a larger data set and by following up the censored observations in the original data set, to record the true survival times is essential. The data set used in the paper does not lend itself to a more complete evaluation as all censored observations have values greater than 60 months and uncensored observations are all less that 60 months. This explains the peculiar statistics for Type 2 and 3 observations with respect to the number of predicted values that are greater than the observed values. One issue that has already presented itself as a clear goal for further research is that some categorical attribute value may not have a defined survival curve as all observations may be censored. How do we compute its distance from other categorical values? Another research issue arising from this work is the need to develop a fitness function for the genetic algorithm that optimises not only the informativenss but also the accuracy of the model.

Acknowledgments

The authors would like to acknowledge funding from the Nuffield Foundation that partially supported the research presented in the paper. We would also like to thank Dr. Ann Smith for her useful medical statistics perspective.

References

[Anand et al., 1996] S. S. Anand, D. A. Bell, J. G. Hughes. A General Framework for Database Mining based on Evidential Theory, *Data and Knowledge Engineering Journal*, 18:189 - 223, 1996.

[Anand and Hughes, 1998] S. S. Anand, J. G. Hughes. Hybrid Data Mining Systems, *Proceedings of the 2nd Pacific Asia Conference on Knowledge Discovery and Data Mining*, pages 13--24, 1998

[Anand et al., 1999] S. S. Anand, A. E. Smith, P. Hamilton, J. S. Anand, J. G. Hughes, P. Bartel. An Evaluation of Intelligent Prognostic Systems for Colorectal Cancer, *Artificial Intelligence in Medicine*, 15(2):193--214, 1999.

[Collett, 1994] D. Collett. *Modelling Survival Data in Medical Research*, Chapman and Hall, 1994.

[Faraggi and Simon, 1995] D. Faraggi, R. Simon. A Neural Network Model for Survival Data, *Statistics in Medicine*, 14:73--82, 1995.

[Kasif et al., 1998] S. Kasif, S. Salzberg, D. Waltz, J. Rachlin, D. Aha. A Probabilistic Framework for Memory-Based Reasoning, Artificial Intelligence 104(1-2):297—312, 1998.

[Lavrač, 1998] Nada Lavrač. Data Mining in Medicine: Selected Techniques and Applications, *Proceedings of the Second International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages. 11--31, London, 1998.

[Wyatt, 1995] J. Wyatt. Nervous about Artificial Neural Networks, The Lancet, 346:1175 - 1177, 1995.

Predicting Surgical Outcome in Temporal Lobe Epilepsy Patients Using a Naïve Bayes Classifier.

Antel SB, M.Sc.[1,3], Li LM, MD[2,3], Cendes F, MD, PhD[2,3], Olivier A, MD, PhD, FRCS(C)[2], Andermann F, MD, FRCP(C)[2], Dubeau F, MD, FRCP(C)[2], Kearney RE, PhD[1], Shinghal R, PhD[4], Arnold DL, MD, FRCP(C)[2,3]

Departments of Biomedical Engineering[1], and Neurology and Neurosurgery[2], McGill University.
MR Spectroscopy Unit[3], Montreal Neurological Institute and Hospital.
Department of Computer Science[4], Concordia University.
Montreal, Canada

Objective: To develop a machine learning-based classifier to predict surgical outcome in patients with temporal lobe epilepsy (TLE).

Method: We studied 81 patients with medically refractory TLE who underwent surgical treatment. In addition to being clinically evaluated, patients were pre-surgically investigated with EEG, proton magnetic resonance spectroscopic imaging (MRSI) and magnetic resonance volumetric (MRV) analysis. Outcome was measured using Engel's classification system, which we adjusted by combining Classes I & II (denoted as Group I) and Classes III & IV (denoted as Group II). A leave-one-out naïve Bayes classifier was developed, using results from the above investigations as inputs.

Results: The naïve Bayes classifier correctly predicted the surgical outcomes of 49/54 (91%) of Group I patients, and 16/27 (59%) of Group II cases. The overall accuracy rate in predicting outcome was 65/81 (80%)

Conclusion: Reliable pre-surgical evaluation of a patient's chances for a successful surgical outcome is feasible using machine learning techniques. Predictive factors can be found among MRSI and MR volumetric data.

Introduction

Surgical treatment, via a selective amygdalohippocampectomy (SAH) or an anterior temporal lobe (ATL) resection, has been shown to be an effective means of seizure control for about 70-80% of patients with medically refractory temporal lobe epilepsy (TLE) [1]. However, pre-surgical investigation of patients is costly, requiring weeks of hospitalization and monitoring via video EEG and telemetry in highly specialized units. Even with such careful observation, approximately 20-30% of patients will not obtain maximal benefit from surgery. In addition, any neurosurgical procedure carries a certain degree of risk. Therefore, given the risks and costs associated with this type of treatment, a more efficient means of pre-operative identification of those TLE patients who stand to benefit from surgical intervention would be a valuable tool. Our aim in this study was to develop a machine learning-based classifier to perform this task.

There are four criteria that such a classifier should fulfill to be of maximal clinical utility. First, it must be able to make a prediction for an individual patient, rather than making generalizations for a group of patients. Secondly, the classifier must use only pre-operative features as inputs. Thirdly, the classifier should be able to provide a measure of how confident it is in each prediction, i.e., posterior probabilities should be measurable. Lastly, the workings of the classifier must be transparent enough that a neurosurgeon or clinical epileptologist without a background in artificial intelligence can understand the "reasoning" upon which a prediction is based, without in-depth knowledge of the underlying mathematics.

A number of studies have examined various factors as they relate to the prediction of surgical outcome in TLE patients. While these studies provide valuable information which can be used when designing a classifier, none fulfill all the above-stated criteria.

Many studies have investigated the relationship between hippocampal atrophy and surgical outcome. A consensus finding is that unilateral hippocampal atrophy is a predictor of a good surgical outcome [2-5]. Bilateral hippocampal atrophy has been reported to reduce the chances for a good outcome [3]. Among patients with bilateral hippocampal atrophy, a recent study from our unit has shown that proton magnetic resonance spectroscopic imaging (MRSI) yields important information for the prediction of surgical outcome. Specifically, a low ratio of N-acetylaspartate (NAA, a marker of neuronal integrity) to creatine (Cr) in the contralateral posterior temporal lobe suggests diminished chances for a good outcome [6]. However, because these studies examine only group differences, they do not make predictions of an individual's chances for a successful surgical outcome.

Several studies have been able to generate individual predictions for surgical outcomes. One study reported success at discriminating between a group of completely seizure free TLE patients and a group of TLE patients who were nearly seizure free following surgery using artificial neural networks [7]. However, it may be difficult for physicians to understand how a neural network produces its predictions, which diminishes its utility in a clinical context.

A recent study [8] reported encouraging success at developing a predictive model, using logistic regression, to discriminate between epilepsy (both TLE and extra-TLE) patients with an excellent chance of being seizure-free post-surgically and those with less than a 50% chance of the

same. However, post-surgical pathological analysis of excised tissue is used to obtain a feature in their model, and thus, while their classifier is a promising prognostic tool, it cannot be used to assist in pre-surgical evaluation of patients.

We have developed a naïve Bayes classifier to predict surgical outcomes of TLE patients, and which meets the conditions set out above. The classifier is based primarily on pre-operative magnetic resonance volumetry (MRV), which allows quantitative measurement of brain structures, and proton magnetic resonance spectroscopic imaging (MRSI), which permits *in vivo* measurement of brain metabolites. Both modalities are non-invasive and require approximately an hour to perform. The classifier is conceptually simple, can produce a prediction for each individual patient, and can attach a posterior probability to each prediction.

Methods

Patients

Our patient database consisted of 81 patients diagnosed with TLE (mean age 35 +/- 11.2 years). The database consisted of 31 males and 50 females. All patients underwent surgical treatment for TLE; 41 patients underwent anterior temporal lobe (ATL) resection, and 40 patients underwent a selective amygdalohippocampectomy (SAH). No significant differences were found between these two patient groups on any of the variables available for this study. Surgical outcomes were assessed using Engel's modified classification scheme [18]. The breakdown of the patients' surgical outcomes was as follows: 53 patients with Class I outcome (free of seizures or residual auras), 1 with Class II outcome (less than 3 seizures per year), 12 with Class III outcome (worthwhile improvement), and 15 with Class IV outcome (no worthwhile improvement). We consolidated the patients into two groups (denoted as Group I and Group II) to obtain larger and somewhat less disparate class sizes. Group I (n=54) consisted of patients who were seizure-free or nearly seizure-free following surgery (Engel's Class I & II). Group II (n=27) consisted of the remaining patients (Engel's Class III & IV).

MR Investigations

MR imaging was performed on a 1.5T scanner (Philips Medical Systems, Best, The Netherlands). Sagittal and coronal T1 weighted images were acquired (TR=550 ms, TE=19 ms), followed by axial proton density (TR=2000 ms, TE=20 ms) and T2 weighted (TR=2100 ms, TE=20, 78 ms). In order to perform volumetric studies of the hippocampi and amygdalae, T1 weighted, 1 mm thick, contiguous slice gradient-echo volume acquisition of the whole brain was acquired. Quantification of volumes was performed using previously published methods [9].

MRSI of the temporal lobes was performed on the same scanner, in a separate session. Following acquisition of scout images in the axial and sagittal planes, a multi-slice transverse spin-echo MRI (TR=2000 ms, TE=30 ms) was obtained. The volume of interest (VOI) incorporated part of the head of the hippocampus, as well as the entire body and tail of the same. Also included in the VOI were portions of gray and white matter in the mid and posterior temporal lobe. The dimensions of the VOI were 85-100 mm on the left-right axis, 75-95 mm on the anterior-posterior axis, and 20 mm in thickness.

A water suppressed MRSI was acquired from the VOI (TR=2000 ms, TE=272 ms, 250x250 mm FOV, 32x32 phase-encoding steps), followed by a MRSI without water suppression (TR=850 ms, TE=272 ms, 250x250 mm FOV, 16x16 phase-encoding steps). Post-processing included zero-filling the water unsuppressed MRSI to obtain 32x32 profiles, followed by the application of a mild Gaussian k-space filter and an inverse 2D Fourier transformation to both the water suppressed and unsuppressed MRSI. The resulting time domain signal was left-shifted and subtracted from itself to improve water suppression [10]. Baseline-correction and measurement of resonance peak areas were performed on the individual spectra using locally developed software.

Fifty-two healthy controls were examined with MR volumetry (1 mm slices: 30 patients, 3 mm slices: 22 patients); MRSI was performed on 51 healthy subjects. MRSI and volumetric data were expressed as Z-scores, which describe the number of standard deviations above or below the mean value of the normal controls.

EEG Investigation

All patients underwent prolonged video-EEG monitoring, using the International 10-20 system including sphenoidal electrodes. EEG data were represented as a label indicating predominant hemisphere(s) of seizure origin: Left or Right (greater than 90% of seizures originating from one side), Left>Right or Right>Left (greater than 70% of seizures originating from one side), or Bilateral (less than 70% of seizures originating from one side). Although EEG data were not used as inputs to the classifier, designation of the various brain structures as ipsilateral or contralateral was made in reference to the EEG results.

Design of naïve Bayes classifier

A naïve Bayes classifier [11,17] is a machine learning technique that assigns an instance consisting of a number of attributes $a_1, a_2,...a_n$ to the most likely class $v_{nb} \in V$, where V is the set of possible outcomes:

$$v_{nb} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, ..., a_n) \qquad (1)$$

Using Bayes' theorem, equation (1) can be expressed as

$$v_{nb} = \arg\max_{v_j \in V} \frac{P(a_1, a_2, ..., a_n \mid v_j) P(v_j)}{P(a_1, a_2, ..., a_n)} \qquad (2)$$

where $P(v_j)$ represents the prior probability of a randomly selected training example having outcome v_j. Since $P(a_1,a_2,\ldots,a_n)$ is a constant independent of outcome group, equation (2) simplifies to

$$v_{nb} = \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n \mid v_j) P(v_j) \quad (3)$$

Class-conditional independence amongst the attributes is assumed, so that equation (3) can be simplified to

$$v_{nb} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \quad (4)$$

We calculated $P(a_i|v_j)$ using the Bayesian approach [11]:

$$P(a_i \mid v_j) = \frac{n_c + mp}{n + m} \quad (5)$$

where n_c=is the number of training examples with a particular value of a_i and outcome v_j; n is the total number of training examples with outcome v_j; p is the prior estimate of $P(a_i|v_j)$; and m is the equivalent sample size. Since all variables fed into the classifier were transformed into binary variables, we set p=½. We chose an equivalent sample size of 2, as we used two outcome groups. Posterior probability for each prediction (i.e., a measure of how confident the classifier was of the individual prediction) can be calculated as

$$\text{posterior probability} = \frac{P(v_{nb}) \prod_i P(a_i \mid v_{nb})}{\sum_j [P(v_j) \prod_i P(a_i \mid v_j)]} \quad (6)$$

Due to the limited number of patients, we elected to use the leave-one-out technique, whereby each of N instances is classified using the other $N-1$ instances as the training set. This technique provides an almost unbiased estimate of the true accuracy of the classifier, serving as a cross-validation of our model [12]. The classifier was implemented in MATLAB 4.2 (The MathWorks Inc., Natick, MA) running on a Red Hat Linux 5.2 platform.

An initial set of attributes was selected for inclusion in the classifier, based on whether a significant difference existed for an attribute across the two outcome groups. Other features were added based on findings in the literature, such as the presence bilateral hippocampal atrophy. Clinical factors such as age and gender were initially included for completeness. Attributes were then added or deleted as needed to increase the accuracy of the classifier. Table 1 summarizes the attributes ultimately used in the classifier. Within the table, v_1 and v_2 refer to Groups I and II, respectively.

Table 1. Inputs to the naïve Bayes classifier and their estimated probabilities.

Attribute	$P(a_i\|v_1)$	$P(a_i\|v_2)$
Sex=Female	0.679	0.500
Unilateral hippocampal atrophy	0.696	0.536
Non-lateralizing bilateral hippocampal atrophy	0.018	0.107
No amygdaloid atrophy & R-score < -0.03	0.161	0.500
Low NAA/Cr in contralateral posterior temporal lobe & contralateral Hc atrophy.	0.071	0.286

Hippocampal atrophy, amygdaloid atrophy, and low NAA/Cr were defined on the basis of a Z-score less than –2. The R-score referred to in the table is defined as NAA/Cr Z score for the contralateral posterior temporal lobe divided by age, a measure which we empirically found to be useful for distinguishing between groups I and II.

It should be noted that some of the parameters in Table 1 do not meet the theoretical requirement of mutual independence. However, it has been shown that a naïve Bayes classifier can, in practice, achieve optimal results even if this assumption is violated [13].

Results

The naïve Bayes classifier developed in this study correctly identified 49/54 (91%) of Group I patients, and 16/27 (59%) of Group II cases. The overall accuracy rate in predicting outcome was 65/81 (80%). Specificity (the number of cases *correctly* predicted to be in a group divided by the *total* number of cases predicted to be in that group) was 82% (49/60) for Group I patients and 76% (16/21) for Group II patients, respectively. Figure 1 displays the accuracy and specificity for Groups I and II. Univariate analysis revealed that NAA/Cr in the contralateral posterior temporal lobe was significantly lower (p<.001) in Group II patients compared to Group I patients.

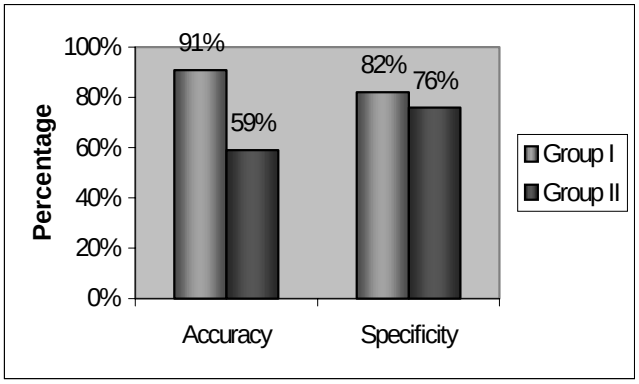

Figure 1. *Accuracy and specificity of classifier.*

Discussion

The naïve Bayes classifier developed in this study provides a simple method of predicting surgical outcome in TLE patients. A key advantage of our classifier is the ability to identify patients who will not be free or nearly free of disabling seizures following surgery. Over one-third of the patients identified as surgical candidates via conventional means (i.e., consensus interpretation by expert epileptologists and neurosurgeons of EEG, neurological, neuropsychological, and neuroradiological investigations) did not experience a complete or near-complete elimination of seizures post-surgically; our classifier identified almost two-thirds of these patients.

However, it is problematic to use the decision of whether to operate (made via conventional means as described above) as a comparison for our classifier. A decision to operate does not necessarily imply an expectation of a seizure-free or nearly seizure-free outcome. The situation does arise whereby a patient will be operated upon in the full knowledge that a Class III or Class IV outcome is all that can be hoped for, based on the logic that a small improvement may be worth the risk to a particular patient. Furthermore, patients deemed unsuitable candidates for surgery did not undergo an operation, and therefore it cannot be determined if the decision not to operate (indicative of an expectation of poor surgical outcome) was correct.

Straightforward interpretation was one of the criteria for a clinically useful classifier set out earlier. The conditional probabilities for each attribute given each outcome class used in the naïve Bayes classifier in essence constitute a frequency table that can provide insight as to how the predictions are made, even without knowledge of the algorithm behind the classifier. The majority of the attributes (3 out of 5) included in the classifier tested for the absence of lateralizing abnormalities; either because the attribute was bilaterally abnormal, or because the attribute was bilaterally normal. Bilateral involvement suggests widespread abnormalities, rather than a focal abnormality than can be easily excised via surgery. Lack of an abnormality suggests an inability to localize the point(s) of seizure origin, complicating the effective resection of the seizure focus.

Previous studies report on the relationship between non-lateralizing hippocampal atrophy and poor surgical outcome [3,14]. We found analogous results in the spectroscopic data. NAA/Cr Z-scores in the contralateral posterior temporal lobe were significantly lowered in Group II patients. This is reflected in two of the features used in the classifier: R-score (NAA/Cr Z-score in the contralateral posterior temporal lobe divided by age) and the combination of a low NAA/Cr Z-score in the contralateral posterior temporal lobe and contralateral hippocampal atrophy.

The only lateralizing feature used in the classifier was the presence of unilateral hippocampal atrophy, which has been widely reported as correlating with a positive surgical outcome [2-5]. An apparently novel finding is that gender is a predictive factor.

Perhaps the most important aspect of the classifier developed in this study is that it does not directly rely on surface or intracranial EEG recordings to make its predictions. Methods of lateralizing seizure focus in TLE patients have been developed [9] that approach the efficacy of EEG monitoring, using only MRSI and MRV data. Thus, a classifier independent of EEG results could eventually lead to an integrated system of lateralization and prognostication, based only on MR data, that would reduce time and hospitalization expenses, and eliminate the risks associated with the implantation of depth electrodes. It should be noted, however, that the contralateral and ipsilateral designations in this and other studies are defined relative to EEG findings, as EEG is the current gold-standard in clinical epileptology. Therefore, a classifier such as the one developed in this study will only be truly EEG-independent upon the emergence of the combined MR investigations as the gold-standard for the evaluation of TLE.

Further refinements include expanding the classifier to classify patients into one of the four outcome classes in Engel's scheme, rather than the two consolidated groups we used. A larger database of patients than is currently available may help facilitate this by providing sufficient sample size for each of the four outcome classes. A larger database would also allow us to detect more subtle differences between the groups, perhaps increasing the accuracy of the classifier.

It is premature to state that a classifier such as ours can make the ultimate decision to operate on a particular patient. Reduction in seizure frequency is only one aspect of surgical outcome; post-surgical cognitive function of the patient is also an important consideration when deciding whether to operate. The expansion of the classifier to include neuropsychological data to help predict such a measure is therefore an important future objective. Furthermore, an outcome other than complete or near-complete elimination of seizures may in fact be a worthwhile improvement for a percentage of TLE patients with particularly frequent seizures. The individual circumstances of patients will still need to be considered when evaluating the surgical option. Nevertheless, the results of this study suggest that our classifier may provide assistance in identifying surgical candidates.

References

1. Engel J. Update on surgical treatment of the epilepsies. *Neurology* 1993; 43(8); 1612-1617.

2. Radhakrishnan K, So EL, Silbert PL, Jack CR, Cascino GD, Sharbrough FW, O'Brien PC. Predictors of outcome of

anterior temporal lobectomy for intractable epilepsy. *Neurology* 1998;51:465-471.

3. Arruda F, Cendes F, Andermann F, Dubeau F, Villemure JG, Jones-Gotman M, Poulin N, Arnold DL, Olivier A.. Mesial atrophy and outcome after amygdalohippocampectomy or temporal lobe removal. *Ann Neurol* 1996;40:446-450.

4. Knowlton RC, Laxer KD, Ende G, Hawkins RA, Wong STC, Matson GB, Rowley HA, Fein G, Weiner MW. Presurgical multimodality neuroimaging in electoencephalographic lateralized temporal lobe epilepsy. *Ann Neurol* 1997;42:829-837.

5. Cascino GD, Trenerry MR, Sharbrough FW, So EL, Marsh WR, Strelow DC. Depth electrode studies in temporal lobe epilepsy: relation to quantitative magnetic resonance imaging and operative outcome. *Epilepsia* 1995;36(3):230-235.

6. Li LM, Cendes F, Antel S, Serles W, Andermann F, Dubeau F, Olivier A, Arnold DL. Prognostic features in surgical outcome of patients with bilateral hippocampal atrophy and intractable temporal lobe epilepsy (Abstr.). *Neurology* 1999; 52 (Suppl. 2): 161-162.

7. Grigsby J, Kramer RE, Schneiders JL, Gates JR, Smith WB. Predicting outcome of anterior temporal lobectomy using simulated neural networks. *Epilepsia* 1998; 39(1):61-66.

8. Berg AT, Walczak T, Hirsch LJ, Spencer SS. Multivariable prediction of seizure outcome one year after resective epilepsy surgery: development of a model with independent validation. *Epilepsy Research* 1998; 29(3): 185-194.

9. Cendes F, Caramanos Z, Andermann F, Dubeau F, Arnold DL. Proton magnetic resonance spectroscopic imaging and magnetic resonance imagin volumetry in the lateralization of temporal lobe epilepsy: a series of 100 patients. *Ann Neurol* 1997;42:737-746.

10. Roth K, Kimber BJ, Feeney J. Data shift accumulation and alternate delay accumulation techniques for overcoming the dynamic range problem. *J Magn Res* 1980; 41;302-309.

11. Mitchell TM. Machine Learning. Boston: WCB/McGraw Hill, 1997.

12. Hand DJ. Discrimination and Classification. Chichester: Wiley, 1981.

13. Domingos P, Pazzani M. Beyond independence: conditions for the optimality of the simple Bayesian classifier. *Proc 13th Int'l Conf Machine Learning* 1996;105-112.

14. Jack CR, Sharbrough FW, Cascino GD, Hirschorn KA, O'Brien PC, Marsh WR. Magnetic resonance image-based hippocampal volumetry: correlation with outcome after temporal lobectomy. *Ann Neurol* 1992; 31:138-146.

15. Bloom D, Jasper H, Rasmussen T. Surgical therapy in patients with temporal lobe seizures and bilateral EEG abnormality. *Epilepsia* 1960; 1; 351-365.

16. So N, Olivier A, Andermann F, Gloor P, Quesney LF. Results of surgical treatment in patients with bitemporal epileptiform abnormalities. *Ann Neurol* 1989; 25(5); 432-439.

17. Duda RO, Hart, PE. Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973.

18. Engel J, Van Ness PC, Rasmussen TB, Ojemann LM. Outcome with respect to epileptic seizures. In: Engel J, ed. Surgical treatment of the epilepsies. 2nd ed. New York: Raven Press, 1993:609-621.

Identification of patients at high cardiovascular risk : a critical appraisal of applicability of statistical risk prediction models.

Hervé Dréau **Isabelle Colombet** **Patrice Degoulet** **Gilles Chatellier**

Medical Informatics Department
Broussais Hospital
96 rue Didot, 75014 Paris.
France

Abstract

Assessment of cardiovascular risk is widely proposed as a basis for taking management decisions among patients presenting with hypertension or hypercholesterolemia. Our aim was to critically assess use of risk equations derived from epidemiological studies in the purpose of identifying high risk patients.

Risk equations were retrieved in the MEDLINE database and then applied to a data set of 118 patients. This data set was an evaluation study of the clinical value of World Health Organization 1993 hypertension guidelines for the decision to treat mild hypertensive patients We calculated agreement: 1) between equations and 2) between equations and the decision to treat taken by the physician.

Most models were not applicable to our population, mainly because the original population had a narrow age range or comprised only males. Between-model agreement was better for the lower and upper risk quintiles than for the 3 other risk quintiles (0.58, 0.33, 0.34, 0.45, 0.70, from the lower to the upper risk quintile). When using an arbitrary threshold for defining high risk patients(i.e. >2% per year), we observed a huge variation of the proportion of patients classified at high risk (from 0 to 17%). There was a poor agreement between risk models and the decision to treat taken by the physician.

These results suggest that risk-based guidelines should be validated before their diffusion.

1. Background

All guidelines related to the cardiovascular field (hypertension, diabetes or lipid management) presently propose to manage patients using an explicit reference to the absolute cardiovascular risk[1]. Many epidemiological studies have provided various risk prediction statistical models in the cardiovascular domain. The characteristics of these studies vary widely in terms of design, origin of the study population, inclusion criteria, measured outcome criteria and period of follow-up. Moreover, the models also differ both in the type of underlying statistical method and in the predictive variables they used.

Therefore, using these models to predict the cardiovascular risk of a given individual, could be questionable. Before choosing a given model, it is mandatory to compare characteristics of the actual population to which the considered individual belongs to those of the original population, to define the internal validity of the model (quality of the study, range of the predictor variables…) and to obtain data on external validity (test of the model in another population).

Among the cardiovascular risk models, those obtained from the Framingham study have been validated in various populations, in the United States, in Australia and in Europe. However, in Europe, and particularly in France, where prevalence of coronary heart disease (CHD) is low[2], several other models are available. Laurier et al addressed the problem of absolute predictive performance and corrected the original Framingham model by calibrating it to the French population[3]. Other models from Germany, Australia, and Scotland could also be used. Ideally, it would be necessary to assess the predictive performance of each model against the real risk. As a first step toward validation, the present work compare how the risk estimated by different models could be predictive of a medical decision taken in hypertensive patients by a physician following a current validated international practice guideline.

2. Objective

To evaluate the usability and the agreement of cardiovascular risk prediction models derived from several large epidemiological studies, in the context of the decision to treat mild hypertension.

We first examined the discriminative performance of the models by looking at the ability of models to classify patients according to levels of risk. Then, performance of each model was assessed by reference to the physician decision to treat or not to treat patients with antihypertensive drugs.

3. Methods

3.1 Data set

We used data from a previously published clinical study [4]. Briefly, non-obese patients referred at the Broussais Hospital hypertension clinic with untreated suspected or known essential uncomplicated mild hypertension, aged 21 years or more were included in a protocol designed for identifying those patients needing treatment. At inclusion, the 118 included patients who had usual laboratory and other diagnostic tests and were then followed up each month, for 6 months. Need for treatment was determined by a physician following the World Health Organization 1993 guidelines for the treatment of mild to moderate hypertension : briefly, drug treatment could be instituted on the basis of both diastolic or systolic blood pressure (BP) level over repeated visits and the physician's estimate of cardiovascular risk according to known risk factors for CHD and stroke. Physician's decision was considered as the gold standard in the present work.

3.2 Models: analysis of the literature

Models of cardiovascular risk were retrieved through a MEDLINE search. A model was selected if 1) it was based on a prospective cohort study; 2) it provided an estimate of absolute risk. A model was defined as *calculable* when it provided all the parameters necessary to calculation, and as *usable* when all necessary variables were available in our data set.

Finally, *usable* models were applied to the 118 patients of our data set, using the data of original papers for defining *applicability* of each model. Thus, minimum and maximum values of each quantitative variables were used as applicability criteria of a given model for a given patient. For example, a model developed on a sample having a diastolic BP between 90 and 120 mm Hg was applied only to patients having a diastolic BP within this range.

3.3 Statistical methods

To assess agreement between the different models, we used a 2-class Kappa coefficient for unbalanced observations based on a tertile classification of absolute risk estimated by each model. High-risk patients were those belonging to the upper tertile, and low-risk patients those in the 2 other tertiles [5]. For models using the same subset of our data set, we used a 5-class Kappa coefficient based on a quintile classification of absolute risk estimated by each model.

Area under the Receiver Operating Characteristics curve (ROC) was used to assess and compare risk classifications derived from the various models to our gold standard, the decision to treat taken by the physician.

The R statistical software (Ross I, Gentleman R. R: A Language for Data Analysis and Graphics *Journal of Computational and Graphical Statistics* 1996;5:299-314) was used for calculations.

4. Results

1.1 4.1 Description of models

According to our inclusion criteria, 27 calculable models from 8 epidemiological studies were identified. Epidemiological characteristics of these models are summarized in Table 1. Three different statistical models were used : Cox regression model (n=4), multiple logistic regression model (n=11) and Weibull time failure accelerated regression model (n=12).

Among the 27 calculable models, 26 were usable. The Laurier equation was not obtained through an epidemiological study, but represents the calibration of a CHD Framingham equation to the French population.

The selected models comprised 4 to 13 different predictors (variables or combination of variables) among which only age was common to all models (BP and tobacco consumption are in all models but with different coding). For a given variable, definition and measurement were not the same across the studies: differences concerned BP (number of measurements), tobacco consumption (quantitative or qualitative variable), left ventricular hypertrophy.

The outcome variable could be either a specific disease (i.e. myocardial infarction), or a composite end-point (i.e. CHD). Definition of outcomes also varied across studies. For example, even in models coming from the same epidemiological study, Framingham, CHD could be described differently (i.e. comprising or not angina pectoris).

4.2 Application to data set

The Framingham and Laurier models could not be used in 4 of our patients (all are aged less than 30 years). With the PROCAM model, only 45 patients remained eligible for calculation (49 were excluded because they were women, and 23 men were excluded because they were outside the age range). The Busselton models could not be used in 24 of our patients (all because of an age < 40 years). The CCHS model could be applied in only 28 patients (all the other were excluded because they were under 55 years). The French Paris Prospective Study (PPS) model could be applied only to 26 patients (all women and 43 men outside the age range were excluded). The Dundee and ERICA models could be applied in 41 (all women and 28 men were excluded). For 4 patients there were no applicable models.

Agreement

Using the 2-class kappa coefficient, the agreement beyond chance between the 26 usable models applied to 114 /118

patients of our data set was 0.68, a value corresponding to a good agreement.

Using the kappa coefficient after classification into quintiles, the Framingham (n=18), and Laurier models applied to 114 patients. They had a global agreement beyond chance of 0.48. Agreement was better for the lower and upper risk quintiles (0.58, 0.33, 0.34, 0.45, 0.70, from the lower to the upper quintile).

Validation against the gold standard

In our data set, 48 of the 118 patients (40.1%; 95% CI: 31.9% - 50.1%) have been treated with antihypertensive drugs at the end of the follow-up. These patients correspond to high risk patients defined by a physician applying a practice guideline. Area under the ROC curve of the various models against the physician's decision (gold standard) ranged from 0.44 to 0.78 for the different models (table 2). When using an absolute cut-point for defining high-risk patients needing a pharmacological treatment, for example the widely used. 2 % annual risk proposed in the guidelines of the European Society of Cardiology, there is a huge variation in the proportion of patients classified at high risk, depending on both the study setting, and the outcome used (table 2). The proportion of patients classified at high risk and their confidence intervals for selected studies were as follows:

- Framingham (CHD), 11 / 114 patients (9.7%; 95% CI: 4.2% - 15.1%)
- Framingham (CVD), 17 / 114 patients (14.9%; 95% CI: 8.4% - 21.4%)
- Laurier (CHD), 6 / 114 patients (5.3%; 95% CI: 1.2% - 9.4%)
- Procam (MI) , 3 / 45 patients (6.7%; 95% CI: 1.7% - 19.3%)
- Busselton (CHD death), 1 / 94 patients (1.1%; 95% CI: 0.06 % - 6.6%)
- PPS (CHD), 0 / 26 patients (0% ; 95% CI: 0.35% - 16.2%

Table 2 also provides the proportion of patients having a risk below 2% per year and 0.5% per year (a very low risk level for which only non pharmacological treatment would be advised). Both proportions varied widely according to studies. However, whatever the model chosen, a high proportion of patients at low risk or even at very low risk were treated with drugs by the physician.

1. Discussion

Many diseases are now preventable if adequate screening and treatment are performed. For example, systematic screening is proposed for colon and breast cancer after 50 years, and at any age for cardiovascular disease. However, to avoid screening in low risk patients, it would be interesting to target interventions or screening towards high risk patients. In this respect, using statistical models based on large prospective studies seems to be the best solution. In the cardiovascular field an important number of potential predictive risk models usable in prevention are available, and the present review is probably not complete. A more exhaustive selection and comparison of logistic risk models can be found in Chambless and al [6].

When trying to find a model, the first obstacle is the lack of information on some parameters of the model in the original paper: often only relative risk for the different risk factors are given. Second, variables found in some models may not be available in the user data set, either for coding reasons (tobacco consumption rather than number of cigarettes per day), or missing variable in the database. Even if the variable is present, definition may differ since there is no agreement on the way of measuring a given factor: for example, how many measurements define the blood pressure? Third, inclusion criteria and range of variables limit applicability to population having comparable characteristics. Thus, even when using the most applicable model (Framingham study) it was not possible to calculate an absolute risk for all the people referred in an hypertension clinic. The relatively low agreement between models result probably from a mixture of measurement, statistical and epidemiological differences.

The practical use of equation is made difficult by the choice of equation and the choice of a risk threshold. The same patient could be classified at high or low risk just because a different equation has been used for calculating risk (Table 2). This is a major problem. In the meanwhile, it seems therefore advisable to favor the use of the Framingham equation, because it is more applicable than the other ones, and has been widely validated outside the United States. The level of absolute risk beyond which define a high risk patient (i.e a patient needing treatment) is purely arbitrary. For example, European and New Zealand guidelines use a 2 % annual risk, whereas the UK guidelines propose a 3% annual absolute risk.

At a first glance, it seems questionable to use the decision to treat taken by a physician as the gold standard to compare the performance of the risk models. However, this choice enlightens the discrepancy between a decision based on multifactorial absolute risk models and practice even if the physician's decision is based on well established guidelines. There are several explanations to the great percentage of our population who has been treated in spite of a low absolute risk. First, our population is younger (mean age: 51 years) than the Framingham one (mean age: 65 years) and it is well known that the use of absolute risk tends to favor intervention for aged patient because age is an important and not modifiable risk factor of cardiovascular diseases, taken into account in all the models, for example there is no patient at high risk of CHD predicted by the Framingham model in the younger half of our population. A solution to

this problem could be to use the marginal absolute risk (risk of the patient minus the risk of a person of the same gender and age without risk factors) to classify people at high level of risk. However, there is no definition of a "high" marginal risk! Second, a 10-year prediction is too short, and the physician may subjectively use a more long term risk estimate.

In conclusion, we show that there are many limitations to the use of prediction models. Individualized life time prediction will undoubtedly improve identification of high risk patients [7]. However, only a randomized controlled trial comparing a risk-based strategy to the traditional strategy will be able to determine if risk prediction improves the quality of care.

2. References

A complete list of references on models is available from request to the authors.

DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach.. *Biometrics* 1988**;44**:837-845.

1. Recommendations of a Task Force of the European Society of Cardiology and the European Resuscitation Council on The Pre-hospital Management of Acute Heart Attacks. *Resuscitation* 1998;**38**(2)**:**73-98.
2. Renaud S, de Lorgeril M. Wine, alcohol, platelets, and the French paradox for coronary heart disease. *Lancet* 1992;**339**(8808)**:**1523-6.
3. Laurier D, Nguyen PC, Cazelles B, Segond P. Estimation of CHD risk in a French working population using a modified Framingham model. The PCV-METRA Group. *J Clin Epidemiol* 1994;**47**(12)**:**1353-64.
4. Chatellier G, Abergel E, Battaglia C, Menard J. Do WHO-ISH guidelines identify high risk mild hypertensive patients?. *Arch Mal Cœur Vaiss* 1998;**91**(8)**:**967-70.
5. Fleiss JL Statistical methods for rates and proportions,2nd ed. J Wiley. p223-224, 1981.
6. Chambless LE, Dobson AJ, Patterson CC, Raines B. On the use of a logistic risk score in predicting risk of coronary heart disease. *Stat Med* 1990;**9**(4)**:**385-96.
7. Lloyd-Jones DM, Larson MG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease . *Lancet* 1999;**353**(9147)**:**89-92.

Table 1: Main characteristics of the calculable cardiovascular risk prediction models.

Study	n° of models	Country	Study Beginning	age (sex[#])	Statistical model[¶]	Variables	Outcome	Usability
Framingham	12	US	1948	30-74 (1)	W	(a)	Multiples[‡] at 4-10 years	114/118
Framingham	1	US	1948	30-74 (1)	L	(b)	CHD at 10 years	114/118
Framingham	1	US	1948	30-74 (1)	C	(e)	Stroke at 1-10 years	114/118
Framingham	1	US	1948	30-74 (1)	L	(f)	Intermittent claudication at 4 years	114/118
Framingham	1	US	1948	30-74 (1)	L		CHD at 8 years	114/118
Framingham	2	US	1948	30-74 (1)	L	(a)	CHD death at 10 years	114/118
Laurier	1	Fr	-	30-74 (1)	C	(a)	CHD at 4-10 years	114/118
Busselton	1	Au	1966	40-74 (1)	L	(c)	CHD death at 10 years	94/118
Busselton	1	Au	1966	40-74 (1)	L	(c) + BMI	CHD death at 10 years	94/118
Procam	1	De	1979	40-65 (2)	L	(d)	MI at 8 years	45/118
Dundee	1	Scot	1984	40-59 (2)	L	(h)	CHD at 5 years	41/118
PPS	1	Fr	1967	43-54 (2)	C	(g)	CHD at 5 years	26/118
Copenhagen	1	DA	1976	55-84 (1)	C	(e)	Stroke at 10 years	28/118
ERICA	1	Europe	1982	40-59 (2)	L	(i)	CHD death at 6 years	41/118
WCGS	1	US	1960	39-59 (2)	L		CHD at 8.5 years	0/118

[¶]: W: Weibull, C: Cox Model, L: Logistic Model.

[‡]: CHD, cardiovascular disease (CVD), myocardial infarction (MI), Stroke, CHD death, CVD death prediction for a 4 to 10 years interval, using either systolic BP or diastolic BP.

#: 1 both, 2 men only

(a) systolic BP or Diastolic BP, age, gender, smoking status (Y/N), diabetes (Y/N), left ventricular hypertrophy (Y/N), total cholesterol (TC) , HDL cholesterol (HDL-C) with interaction between variable.

(b) five levels for TC, five levels for HDL-C, five levels for BP, diabetes, smoking status, age.

(c) age, Systolic BP, TC, Smoking status.

(d) systolic BP, age, TC, HDL-C, triglycerides, smoking status, familial history of myocardial infarction, angina pectoris.

(e) age, Systolic BP, diabetes (Y/N), smoking status (Y/N), left ventricular hypertrophy (Y/N), atrial fibrillation, CVD, anti-hypertensive treatment.

(f) four levels for both systolic and diastolic BP, gender, tobacco consumption (cig/d), diabetes (Y/N), age, TC.

(g) TC, Systolic BP, tobacco consumption (cig/d), diabetes (Y/N), age.

(h) systolic BP, tobacco consumption (cig/d), age, TC

(i) systolic BP, tobacco consumption (Y/N), age, TC, body mass index (BMI).

Table 2: Influence of model and threshold of absolute risk (AR) on the decision to treat.

Model	n° of patients [¶] (treated/total)	Outcome[‡]	Risk > 2% /years	Untreated and AR > 2%/years	Treated and AR < 2% /years	Treated and AR < 0.5%/years	ROC curve
Framingham (SBP)	48/114	CHD	11	5	42	24/56	0.58
Framingham (DBP)	48/114	CHD	12	5	41	20/54	0.59
Framingham (categorical)	48/114	CHD	10	5	43	14/38	0.56
Laurier Model	48/114	CHD	6	1	43	28/75	0.58
Framingham (SBP)	48/114	CVD	17	8	39	7/30	0.59
Framingham (Cox model)	48/114	stroke	0	-	-	37/94	0.59
Framingham (SBP)	48/114	IC (4)	0	-	-	45/111	0.60
Framingham (SBP)	48/114	CHD (8)	9	4	43	24/37	0.60
Busselton	45/94	CHD-death	2	0	43	39/77	0.51
PROCAM	24/45	MI	3	2	23	18/31	0.44
PPS	13/26	CHD (5)	0	-	-	9/21	0.54
CCHS	9/28	stroke	2	0	7	1/12	0.78
Dundee	23/41	CHD (5)	2	0	21	6/10	0.49
ERICA	23/41	CHD-death (6)	0	-	-	21/39	0.52

[¶]: Number of patients in whom the given equation was applicable.

[‡]: CHD: coronary heart disease, CVD: cardiovascular disease, MI: Myocardial infarction, IC : intermittent claudication.

Influential Case Detection in Medical Prognosis

Lucila Ohno-Machado
Decision Systems Group
Brigham and Women's Hospital and
Health Sciences and Technology Division,
Harvard Medical School/MIT
Boston, U. S. A.

Staal Vinterbo
Knowledge Systems Group
Dept. Computer and Information Sciences,
Norwegian University
of Science and Technology
Trondheim, Norway

Abstract

We have compared different regression diagnostic methods for detection of influential cases in a model that predicts death from traumatic injuries. For the purposes of this work, we defined "influential" as an outlier whose removal might be beneficial to the prognostic model. The regression diagnostic methods involved "unicase" and "multicase" determination of case influence on model predictive performance. Multicase determinations were performed using two methods: a sequential "backward" selection of cases and a non-sequential genetic algorithm.

A case whose removal resulted in a model that had better fit (measured by the area under the ROC curve, or AUC) was considered influential. The unicase and sequential backward selections resulted in a final model that had excellent fit (AUC=0.98 in training set), but that lost significant predictive capability (AUC=0.78 in test set). The genetic algorithm produced a final model that kept fewer cases, had good fit (AUC=0.95), and retained predictive capability (AUC=0.86). These results indicate that a genetic algorithm approach to case selection may yield better results than a unicase or a sequential multicase approach, possibly because of its ability to detect sets of cases that are influential *en bloc*, but may not be sufficiently influential when considered in isolation.

1 Introduction

Prognostic models of trauma have been extensively investigated using different regression and machine learning algorithms and some are currently used in clinical practice [1-4]. Determining case influence is important for building and continuously updating these models as new cases are added. Detecting and removing influential cases that bias prognostic models are desirable in order to evolve a training set of "good" cases. Case influence determination is developed under the name *regression diagnostics*, and involves the identification of not only undesirable outliers, but also cases that have a stronger impact on the prognostic model's estimated parameters or final fit. These features are often highly interrelated.

Considerable debate has centered on what constitutes a good method for influential case detection. Several indices have been proposed: (standardized and studentized) residuals, leverage, Cook's statistic and its variants, etc.[5,6]. Graphical methods to help visualize influence have also been proposed [7]. Other methods have been proposed by machine learning researchers [8,9]. Most influence detection methods are based on the effects of case deletion on the parameters of the model or on its fit and predictive ability. Because of computational intractability, the methods are usually used in a "unicase" manner (i.e., just one case is deleted at a time). However, it has been noted that some cases may not be influential when considered independently, but may become influential when considered *en bloc* [5,6,7]. This influence may be detected if deletion is used in a "multicase" manner (i.e., several cases are deleted at a time). Considering all subsets of cases for influence detection is intractable (requiring 2^m calculations for m cases). Therefore, certain heuristics need to be used in practice.

In this work, we were interested in the detection and removal of undesirable outliers from a particular data set for the construction of a prognostic model that generalized well to a set of previously unseen cases. We have implemented unicase and multicase influence detection methods for logistic regression. Two multicase models were built: one based on backward selection and another based on a genetic algorithm. We describe and compare these different methods and illustrate them with a model aimed at building a prognostic index from a data set of acute trauma patients.

2 Material and Methods

2.1 Data

We used a dataset of 300 patients from a random selection of patients admitted to the University of New Mexico Trauma Center between 1991 and 1994. The complete data set of 300 patients is described in [10] and available on the Internet at http://stat.unm.edu/~fletcher. The collection has been used previously in [11]. Each case in the data set had information on age, Injury Severity Score (ISS), Revised Trauma Score (RTS), Type of Injuries (TI), and outcome (survival or death). The variable ISS has numeric values on a scoring scale, RTS

is continuous and TI is binary ("blunt" or "penetrating"). We selected this data set because we knew from the literature [10] that it could produce reasonable prognostic indices and because its size (in terms of variables and cases) was adequate for this experiment.

For some experiments, the full data set was randomly split into a training (n = 152) and a test (n = 148) set, such that the same number of deaths appeared in each data set. The training set was used to build the models and select cases, and the test set was left out for evaluation purposes. In other experiments, we used the full data set. Table 1 shows the means of each variable for the training, test, and full data sets.

	ISS	TI	RTS	age	survival
Training	14.296	0.243	7.265	31.664	0.927
Test	14.263	0.256	7.309	31.148	0.925
All	14.280	0.25	7.286	31.41	0.926

Table 1. Variable means for training (n = 152), test (n = 148), and full data sets (n = 300).

2.2 Methods

We built a logistic regression model using all cases and refer to it as the "baseline" model. We built unicase and multicase deletion methods using C and the SAS macro language. All logistic regression models were built using the SAS procedure LOGISTIC [12] and the same default parameters. The area under the ROC curve (AUC) was measured for all methods described below.

Unicase deletion

We constructed 152 different training sets in which one patient was removed. The resulting AUC was recorded for each model. The AUCs corresponding to the use of the models in the test set of 148 patients were also recorded for evaluation purposes. Cases were considered influential if the models in which they were removed resulted in high AUCs (as measured in the training set of remaining 151 cases). The AUCs for all models were sorted in descending order. The "most influential" cases were defined as those corresponding to models in the top of the sorted file. For example, the topmost individual case is influential because its removal causes the AUC to increase from 0.91 to 0.95[1].

Multicase deletion

Exhaustive evaluation of all 2^{152} subsets of cases is not feasible. Therefore, we used certain heuristics to construct subsets that were based on sequential stepwise deletion and guided sampling using a genetic algorithm.

[1] The bottom most individual could also have been considered "influential" because its deletion causes the AUC to drop from 0.91 to 0.90, but we have not included this case in our definition.

- *Backward selection*

We used a heuristic borrowed directly from variable selection methods. We start with the full data set, and build models in which each case is removed, as in the unicase selection above. We then remove the case corresponding to the model with highest AUC and use the reduced data set to determine the next case to be removed, using the same criterion. This strategy has been suggested by Cook and Weisberg [6] and Belsey et al.[5], and requires that we construct the following number of models:

$$\prod_{i=0}^{m-1}(n-i)$$

where m represents the number of cases removed.

- *Genetic Algorithm*

For an introduction to genetic algorithms, see [13]. We used a measure of fitness for a selection of cases that has been used in another study for variable selection [14]. Given a training set C, and a selection of cases v, we construct a logistic regression model $l_C(v)$. We evaluate the model using the AUC, and represent this evaluation as $a(l_C(v))$. For a total of n cases, and m cases in selection v, we use the following fitness function:

$$f(v,C) = a(l_C(v)) + r*(n - m)/n.$$

The second term rewards a parsimonious selection of cases, and is weighted by parameter r. We used r = 0 and r =.05 in this experiment. Population size was fixed at 50 (each "individual" in the population is a certain subset of all cases), and cross-over, mutation, and inversion probabilities were set at 0.3, 0.05, and 0.1, respectively. The stop criterion was the lack of improvement of the average fitness of the population after 10 generations.

3 Results

We used the test set in order to assess how generalizable the results would be in a previously unknown set of cases. Table 1 shows that the data split we used preserved the main characteristics of the data set. Since we noted significant differences between the performance on the training and test sets after performing case deletion, we also tested the different case selection methods using the full data set, so that we could identify whether certain cases deemed influential were just artifacts of our particular data split.

3.1 Baseline Regression Diagnostics

Data split (152 training, 148 test)

All cases in the training set were included in this model. The AUC was 0.91 in the training set and 0.88 in the test set. By using the option INFLUENCE in SAS/LOGISTIC [12], we obtained the following 5 most influential[2] cases (in descending order), using the statistics below:

[2] Note that this definition of an influential case does not assume that a case is necessarily a "bad case", as we have defined in this work.

- Pearson residual: 34,120,136,6,24.
- Deviance Residual: 34,120,136,6,90.
- Hat Matrix Diagonal: 4,3,31,46,90.
- C: 4,3,24,120,90.
- CBAR: 4,3,24,120,136.
- Change in Deviance to Deletion: 34,120,136,4,24.
- Change in Pearson χ^2 statistic due to deletion: 34,136,120,24,6.

Full data set
All 300 cases in the training set were included in this model. The AUC was 0.90. Cases from what was the test set in the previous experiment were numbered 153 and higher. The bolded numbers refer to influential cases that coincided with those of the training set:

- Pearson residual: **34**,191,219,**136**,180,**24**,**6**,238,273,249,90.
- Deviance Residual: **34**,191,219,**136**,180,**6**,24,238,**90**,273,249.
- Hat Matrix Diagonal: **4**,83,**3**,129,296,96,139,32,170,211,154,**90**.
- C: **120**,**90**,**24**,139,273,249,191,34,219,238,228,11.
- CBAR: **120**,**24**,90,139,191,273,249,34,219,238,228,117.
- Change in Deviance to Deletion: **34**,191,**120**,219,**136**,**24**,238,90,273,249,11,228.
- Change in Pearson χ^2 statistic due to deletion: **34**,191,**120**,219,**136**,180,**6**,**24**,238,273,249,139.

Note that the cases selected were almost the same as those selected for the data split, with additional cases selected mostly from what constituted the test cases in the previous experiment. This indicates that the data split was probably representative of the full data set.

3.2 Unicase selection

Data split (152 training, 148 test)
Of the 152 models evaluated, the ten models which resulted in the highest AUCs for the respective training sets were those in which cases 34, 24, 120, 136, 83, 69, 29, 45, and 73, were removed, in this order. When one case was removed at a time, AUCs for the 152 training sets ranged from 0.89 to 0.95. For the corresponding models, AUCs for the test sets ranged from 0.80 to 0.89.

We tried to eliminate multiple cases using the unicase selection criterion (e.g., eliminating cases 34, 24, 120, and 136 in the same model). We tried all subsets of size 2 to 9. The maximum number of cases that could be removed without reaching an AUC of 1 (higher AUCs would not be obtained by further removal) for the training set was 9. Removal of 6 cases resulted in an AUC for the training set of 0.99 and an AUC for the test set of 0.77. Removal of 9 cases resulted in an AUC for the training set of 0.99 and an AUC for the test set of 0.62. The difference in performance against that of the baseline model was significant for $\alpha = 0.05$. There was clear overfitting of the data, indicating that the models in which the influential cases were deleted did not generalize well.

Full data set
The results using a split of 152 and 148 cases raised the questions as to whether some of the removed cases were wrongly considered influential ("bad cases") for the model built using the training set, but were actually necessary to build a model that would generalize to the test set. The rationale is that, although the case did not fit the model well, a similar case existed in the test set, so that removing the case was detrimental to performance in the test set. If that were true, the selection of a case as influential would be a pure artifact of the data split. If we used the full data set and still the same variables were selected, however, we could refute this argument.

We built another set of models using 299 cases each (one case removed at a time). We wanted to check whether the same cases would be considered influential. The twelve cases first removed were **34**,191,**120**, 180, 4, **136**, 273, 6, 90, **83**, 32, and **69**. AUC for the training set after removal of 9 cases was 0.98. There was reasonable agreement between the data split and the full data set unicase selection.

3.2 Backward selection

In this method, the case considered most influential is deleted first, then models with *n*-1 cases are reanalyzed in order to select the second most influential case (taking into account the one already removed), and so on.

Data split (152 training, 148 test)
We evaluated 897 different models. Cases 34, 136, 120, 24, 83, 90, and 6 were removed, in this order. At the removal of the 7th case, the AUC in the training set was 1, so no further cycles were performed. The AUCs in the training set after the 6th deletion was 0.99, and that in the test set was 0.78. There was significant difference in performance when compared to the baseline model, for $\alpha = 0.05$. The resulting model did not generalize well.

There was clearly overfitting, since the results in the test set were significantly worse than those of the training set. This method seemed to do no better than the unicase selection method. The five cases deemed most influential by both methods were actually the same: 34, 136, 120, 24, 83, with a slight change in order.

Full data set
In order to verify whether the selection performed by the backward method was an artifact of the data split, we evaluated the method using the full data set. Cases **34**, 191, **136**, 180, 219, **6**, **120**, 273, 69, 234, and 129 were removed, in this order. Again, there seemed to be agreement between the selection from the data split and the full data set. After the 7th deletion, the AUC for the training set of was 0.98. We evaluated 3,534 models.

3.3 Genetic Algorithm

As described before, the selection of cases was based on a fitness function that maximized the AUC in the training set. We expected that influential cases would be removed at each generation (their removal would cause the AUC to increase, as in the previous selection methods), so that only the fittest, non-influential individuals would stay in the training set. The main difference in using this method is that cases would be *considered en bloc*, rather than in a unicase manner. The genetic algorithm selection is also different from backward selection because it does not depend on a particular sequence of removals that is initiated by a unicase selection at each cycle of elimination, as in the backward selection case. The addition of a term that rewarded more parsimonious training sets (or larger removal of cases) did not make the experiments nonparallel to those of the other methods. We could have, for example, obtained the same effect in those methods by just continuing to do influential case elimination until a certain number of cases was left in the training set.

Data split (152 training, 148 test)

Using *r = 0*: The genetic algorithm performed 589 fitness function evaluations before reaching the stop criterion. Eighty-eight cases were left in the model (64 were removed). The AUC in the training set was 0.95 and in the test set was 0.86. The difference in performance when compared to the baseline model was not significant for $\alpha = 0.05$. Cases 34 and 24 were removed (they were also removed in the unicase or backward selection methods), but cases 120, 136, 83, 69, 90, 45, 29, 73, and 6 (which were removed in unicase or backward selection procedures) persisted in the training set.

Using *r = 0.05:* The genetic algorithm performed 669 fitness function evaluations before reaching the stop criterion. Clearly, this number was much smaller than 2^{152}. Curiously, it was also smaller than the one resulting from the backward selection method. Fifty-six cases were left in the model. The AUC in the training set was 0.92 and in the test set was 0.89. The difference in performance for the test set, when compared to the baseline model, was not significant for $\alpha = 0.05$. The reduced model performed as well as the baseline one on the test set and, once defined, could be calculated in about half the time. Cases 34, 120, 69, 45, 90, and 73 were removed (they were also removed in the unicase or backward selection), but cases 24, 136, 83, 29, and 6 (which were removed in one or both of those selection procedures) were not.

Although the results of the genetic algorithm that used the penalty term were slightly better, they were not statistically different from those of the experiment that did not use that term. Of note is the fact that the classification performance of the genetic algorithm on the test set did not degrade as in the unicase or backward selection methods.

Full data set

Using $r = 0$, the genetic algorithm performed 519 fitness function evaluations, leaving 162 cases in the training set. The AUC was 0.90, and cases 120,180,4,136,90, and 29 were removed, among several others. Using $r = 0.05$, the algorithm performed 408 evaluations, leaving 136 cases in the training set. The AUC was also 0.90. Cases 120,136,90,83,73, and 29 were removed, among others. There was some agreement between the cases selected in the data split and the full data set experiments (79 out of 152 were either removed or kept in the training set for both experiments). The fact that the agreement was not higher suggests that considering cases *en bloc* may indeed affect influential case determination.

4 Discussion

The cases considered most influential for the unicase (standard regression statistics and difference in AUC) and backward selection methods were almost the same. This is no surprise, since the backward selection case can be considered an extension of the unicase method to several cycles. As discussed before, these heuristics for case selection are not only intuitively appealing, but also computationally tractable. However, they do not seem to correctly select and remove undesirable outliers, as can be demonstrated by a significant decrease in the classification performance in the test set (AUCs around 0.78 after 6 deletions). The genetic algorithm, however, seemed to correctly identify sets of cases that could be deleted without compromising the generalization of the model (AUCs around 0.86), even after a significant reduction in the number of cases. Different cases were considered influential by the genetic algorithm, and this may have happened because of different coverage of the space of possible solutions. The genetic algorithm selection considers several cases *en bloc*, and is not greedy as is the backward selection method.

In the genetic algorithm experiment, the fact that the AUCs in the test set did not increase after the removal of what we identified as influential cases might indicate that these cases could be considered redundant, instead of "influential". That is, the removal of these cases did result in a model that had better fit to the training data (AUCs around 0.98), but did not decrease nor increase its generalization ability significantly (AUC = 0.86). This may be a feature of this particular data set.

We have used the AUC as a measure of a model's classification performance. Other methods for assessing model fit and predictive classification could have been used, but since we anticipated that the eventual evaluation of these dichotomous outcome prognostic models would use AUC, we concluded that it might be a good indicator of case influence. As we can see by comparing the results of our unicase selection and the ones based on changes in deviance and in the Pearson χ^2 statistic due to deletion, the cases selected using AUC were approximately the same as those using these standard regression diagnostics

monitoring statistics. We have assumed that the case removal resulting in the highest AUCs using the training set was the most influential. One could argue that the case removal resulting in the lowest AUC can also be the most influential (i.e., influence determination could be based on the non-signed *difference* in AUCs from the baseline and the reduced models). The methods presented can be easily adapted to accommodate this change.

As we could observe in retrospect by inspecting the AUCs on the test sets, not all cases considered influential by the unicase and backward procedures were real undesirable outliers. The removal of certain cases adversely affected the performance on prospective cases, indicating overfitting of the model. The fact that the performance of the genetic algorithm selection method was approximately the same as that of the baseline model seems to indicate that the method was successful in eliminating influential cases, while keeping the "good" cases.

Classical influence diagnostics could have been used for case selection for training sets, but a method would need to be developed to determine whether the case was an undesirable outlier or a relatively rare, but representative case. In this experiment, we have shown that simple unicase or sequential selection of cases may not always work. Further experiments using larger data sets are necessary.

Although we have shown an illustration of different methods for detecting influential cases in a particular model of logistic regression, these results may be applicable to other machine learning models as well. We are currently investigating this issue. We cannot ignore the fact, however, that the complexity of certain machine learning models would be added to these already computer-intensive influence detection techniques.

5 Summary and Conclusions

Identification of influential cases in data sets used for machine learning is not a trivial task. Several indices of influence for a particular case have been proposed for logistic regression models. These indices are often used considering just one case at a time (i.e., indices are based on the influence of each case with respect to the data set that uses all n or n-1 cases). Results produced in this "unicase" manner are different from those resulting from a "multicase" detection (indeed a case may not be influential by itself, but it may become so when considered *en bloc*). Prognostic models in medicine need continuous updates as new data are added to the training set. Determining which cases in the training set are influential can facilitate this task. As an illustration of the differences between unicase and multicase influence detection, we used a set of patients with acute trauma and modeled probability of death using a simple logistic regression model. We constructed two variants of the multicase method to select influential subsets: one using backward selection and one using a genetic algorithm. We showed that, by using the genetic algorithm, we obtained a different set of cases that could be considered influential, and it was possible to reduce the training data set significantly without affecting discriminatory performance of the prognostic model.

Acknowledgments

This work was supported in part by grant R29 LM06538-01 and contract 467-MZ-802289 from the National Library of Medicine, and by project grant 107409/320 from the Norwegian Research Council. We thank Stephan Dreiseitl, Robert Greenes, and Aziz Boxwala for reviewing of this manuscript.

References

1. Braithwaite IJ, Boot DA, Patterson M, Robinson A. Disability after severe injury: five year follow up of a large cohort. *Injury* 1998 Jan;29(1):55-9.
2. Clark DE, Ryan LM Modeling injury outcomes using time-to-event methods. *J Trauma* 1997 Jun;42(6):1129-34.
3. Wyatt JP, Beard D, Busuttil A . Quantifying injury and predicting outcome after trauma. *Forensic Sci Int* 1998 Jul 6;95(1):57-66.
4. Osler T, Baker SP, Long W. A modification of the injury severity score that both improves accuracy and simplifies scoring. *J Trauma* 1997 Dec;43(6):922-5.
5. Belsley DA; Kuh E; and Weilsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley and Sons, New York, pp.292, 1980.
6. Cook RD and Weisberg S. *Residuals and Influence in Regression.* Chapman and Hall, New York, pp.230, 1982.
7. Brodley CE and Friedl M.A. Identifying and eliminating mislabeled training instances. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence:799--805. AAAI Press, 1996.
8. Gamberger D, Lavrac N, Dzeroski, S. Noise elimination in inductive concept learning: A case study in medical diagnosis. In: Proc. of the 7th International Workshop on Algorithmic Learning Theory:199-212. Springer, Berlin, 1996.
9. Atkinson AC. *Plots, Transformations, and Regression.* Clarendon Press, Oxford, 1985.
10. Christensen R. *Log-Linear Models and Logistic Regression.* New York: Springer-Verlag, 1997.
11. Bedrick EJ, Christensen R, Johnson W. Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician* 1997, 51, 211-218.
12. The LOGISTIC Procedure. In: *SAS/STAT User's Guide*. SAS Institute, Cary, 1990.
13. Mitchell M. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, 1996.
14. Vinterbo S, Ohno-Machado L. *A Genetic Algorithm to Select Variables in Logistic Regression: Example in Myocardial Infarction*. DSG Technical Report, January 1999.

A specialised POMDP form and algorithm for clinical patient management

Niels Peek
Department of Computer Science, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
E-mail: niels@cs.uu.nl

Abstract

Partially observable Markov decision processes (POMDPs) have recently been suggested as a suitable model to formalising the planning of clinical patient management over a prolonged period of time. However, practical application of POMDP models is hampered by the computational complexity of associated solution methods. It is argued that the full generality of POMDPs is not needed to support many decision problems in clinical patient management, and that specialised forms are often sufficient. A specialised form of POMDP, tailored to a particular type of management problem, is introduced. It is described how a new solution method, based on Monte Carlo simulations of the decision process, can take advantage of this specialised form.

1 Introduction

Managing patients that suffer from a progressive disease is a complicated task involving a mixture of test planning, treatment selection, and prognostic assessment. The large number of possible management strategies over time precludes formalisation of this task using traditional representations such as decision trees and influence diagrams. Recently, partially observable Markov decision processes (POMDPs) [4; 8] have been suggested as a providing a suitable, integrated approach to this type of management problem [7; 10]. POMDPs are models for sequential decision making under conditions of uncertainty and limited observation opportunities. By taking into account both immediate and longterm consequences of decisions, POMDPs provide a powerful framework for decision-theoretic planning of clinical actions. Unfortunately, the computational burden associated with solving POMDPs is overwhelming, precluding their application to problems of practical size [9].

However, for many specialised problems, the full-blown generality of the POMDP approach and its associated solution methods is superfluous. We believe that this holds in particular for clinical decision problems, where often the class of admissible solutions is significantly constrained. In this paper, we discuss a specialisation of POMDPs that is tailored to a frequently re-occurring type of clinical management problem, and propose a solution method that is able to exploit the properties of this specialised form. The management problem we envision to support looks as follows. A patient suffers from a disease from which natural recovery is possible, but which may also cause harmful complications over time. There are possibilities to halt progress of the disease and its complications by intervention (e.g. surgery), but these involve a serious risk to the patient. The main problem is therefore deciding whether or not to intervene, and if so, when. Prior to intervention, it is possible to perform several diagnostic procedures; these procedures reveal information on the clinical state of the patient at the time the procedure is undertaken, but they also comprise a (smaller) risk. A secondary problem is therefore the selection and timing of diagnostic procedures.

2 Model form

In this section, we briefly describe the general POMDP model and its associated solution form. Given a set X of variables, let Ω_X denote the set of all *configurations* of X, i.e. all possible value assignments to variables from X. A POMDP model is a tuple (T, X, A, P, o, L), where

- T is a linearly ordered set of *decision moments*,
- X is a finite set of *stochastic variables*, jointly defining the set Ω_X of *states*,
- A is a finite set of available *actions*,
- $P = \{p_t^a : \Omega_X \times \Omega_X \rightarrow [0,1] \mid a \in A, t \in T\}$ is a set of time- and action-dependent *transition probability functions*,
- $o : A \rightarrow \wp(X)$ is an *observation function*, and
- $L : \{l_t : \Omega_X \times A \rightarrow \mathbb{R} \mid t \in T\}$ is a set of time-dependent *loss functions*.

The set T of decision moments denotes the points in time where the decision maker is expected to select an

action $a \in A$. We restrict ourselves to *finite-horizon problems*, and take $T = \{0, 1, 2, \ldots, N\} \subset \mathbb{N}$. No action is selected at the last decision moment $t = N$; this moment is included for evaluation of the final state only. The clinical state of the patient is described by the set X of discrete, stochastic variables; let $\mathcal{S} = \Omega_X \times \cdots \times \Omega_X = \Omega_X^{N+1}$ denote the set of all possible state sequences. When configuration $\mathrm{c} \in \Omega_X$ characterises the state at time point $t \in T$, selection of action $a \in A$ will result in a transition to state c' at time point $t + 1$ with probability $p_t^a(\mathrm{c}, \mathrm{c}')$. Furthermore, the decision maker is able to observe the configuration of the set $o(a) \subseteq X$ at time point t, and can use this observation to optimise subsequent decision making; note that the observation function o is independent of time. At each decision moment the decision maker also incurs a loss $l_t(\mathrm{c}, a)$; the losses associated with subsequent moments in a realisation of the decision process are combined by a *utility function* $u : \mathbb{R}^{N+1} \to \mathbb{R}$.

Now, let ϕ be a joint probability distribution on X at the initial time point $t = 0$, reflecting the decision maker's prior beliefs on the clinical state of the patient. Given ϕ and a sequence α of action choices for all decision moments (except $t = N$), we obtain a probability distribution on $\Pr_{\phi,\alpha}$ on the set $\mathcal{S}$ of possible state sequences. From this distribution, we can compute the expected utility of action sequence α under ϕ. Our objective is to select actions during the decision process such that expected utility is maximised. Prior to the first action choice, we therefore compose a *decision-theoretic plan* π, which prescribes an action choice for each time point $t < N$, given the history of past actions and observations. When m is the maximum number of distinct observations that may follow an action choice (i.e., if $Y = o(a)$ then $|\Omega_Y| \leq m$, for each action $a \in A$), we have that m^N is an upper bound on the size decision-theoretic plans. The number of possible plans is bounded by k^{mN}, where $k = |A|$ is the number of available actions. It is therefore not surprising that the problem of finding the optimal plan is PSPACE-complete [9].

A POMDP model was recently developed to support the clinical management of patients with *ventricular septal defect* (VSD), a frequently occurring congenital heart disease [10]. A VSD is an abnormal opening in the heart causing heart failure and associated symptoms such as shortness of breath, feeding problems, and growth retardation. Approximately 70% of all VSDs close spontaneously in the first years of life due to tissue growth, obviating the need for surgical intervention. However, in the long run the disease may cause irreversible damage to the lungs and a severely impaired respiratory function. Several diagnostic tests (ECG, echocardiography, cardiac catheterisation, chest X-ray, and pulmonary biopsy) are available to examine the patient's condition before deciding upon cardiac surgery. The cardiologist treating a VSD patient therefore faces the type of management problem described in the previous section. The POMDP model for VSD has 33 state variables, yielding approximately $9,7 \cdot 10^{15}$ possible configurations (i.e. states of the POMDP); many configurations, however, cannot occur in practice, or can only occur in specific circumstances. This is expressed in the transition probability functions by assigning zero probability to those configurations. The model distinguishes 6 decision moments (ages of the patient, ranging from 3 months to 8 years), and 7 distinct actions to choose from.

It was also shown in [10] how *temporal probabilistic networks* can be used to graphically represent the transition probability functions of a POMDP model, and how this representation, by exploiting conditional independence relations between state variables, can strongly reduce the number of probability estimates required to complete the model. This is especially useful when the number of state variables is large: the complexity of transition probability functions quickly grows in the number of variables. A compact representation, as in probabilistic networks, is then indispensable as obtaining probability estimates is often a cumbersome task. We do not further elaborate on this representation here, and refer the interested reader to the paper in question for more details.

3 A specialised POMDP form

We will now propose a special form of POMDP model that is tailored to support the management problem described in Section 1. We first characterise the types of loss and utility function that are used within this special form, and then describe three restricting assumptions we make on actions, transition probabilities, and plan structure.

We take a loss $l_t(\mathrm{c}, a)$, $t < N$, to represent the *mortality risk* associated with state c and action a at time point t, and a loss $l_N(\mathrm{c}')$ to denote *life expectancy* (in years) associated with final state c' at time point $t = N$, where no action choice is made. Let $r_0, \ldots, r_{N-1}$ be such mortality risks, obtained from a given evolution of the decision process (i.e. states and actions for each of the decision moments up to time point $t = N - 1$). Then,

$$s_t = \prod_{i=0}^{t-1} (1 - r_i) \tag{1}$$

denotes the chance that the patient survives at least up to time point $t > 0$. Now, let d_t be the (fixed) actual duration (in years) between the start of the decision process and decision moment t, $0 \leq t \leq N$; we then have that

$$le_t = \sum_{j=1}^{t-1} d_j r_j s_j \tag{2}$$

is the life expectancy of the patient up to time point t. The following utility function u now expresses overall

life expectancy:

$$u(r_0, \ldots, r_N) \quad = \quad le_N + (d_N + r_N) \cdot s_N, \qquad (3)$$

where $r_N = l_N(\mathrm{c}')$ denotes life expectancy at the final time point. This type of utility function is generally referred to as *risk-sensitive* [5]. We note that it is also possible to encode mortality risks in the transition probability functions, but we deliberately choose not to do so, for reasons explained shortly.

We make three further assumptions on the POMDP model and its admissible solutions. First, the set A is taken to be composed of three disjoint sets A_{test}, A_{treat}, and A_{skip}, where A_{test} constitutes the set of available diagnostic procedures, A_{treat} lists treatment alternatives, and A_{skip} is a singleton set that consists of the special action *skip* (i.e. refrain from acting at the specified point in time) only. The set A_{treat} is assumed to be relatively small compared to A_{test}; e.g., in the VSD domain, we have $A_{\text{treat}} = \{surgery\}$ and $A_{\text{test}} = \{ECG, echo, catheter, X\text{–}ray, biopsy\}$. Second, from A_{treat} an action is selected at most once, and after that moment, further action is refrained from (by selecting *skip* for all subsequent moments). Before the moment of treatment though, actions may be selected freely from A_{test} and A_{skip}. From the first two assumptions we thus obtain a restricted set Π of admissible plans, in each of which there is but a single moment of control, preceded by multiple moments of observation. The size of the set Π is bounded by $(k_{\text{test}} + 1)^{mN}$, where $k_{\text{test}} = |A_{\text{test}}|$, and as before, m is the maximum number of distinct observations that may follow an action choice. Although $k_{\text{test}} < k$ (where $k = |A|$), this number of admissible plans is still very large. The *average* size of plans in Π, however, equals $m^{N/2}$.

The third and last assumption is that state development is independent of test actions. So, $p_t^a = p_t^{skip}$ for each $a \in A_{\text{test}}$, $t = 0, \ldots, N-1$. Note that we can make this assumption because mortality risks are encoded in the loss functions: this enables us let all diagnostic procedures induce the same transition probabilities, even if they differ with respect to their associated risks. Without this assumption, each of the k^N possible action sequences α induces a different probability distribution $\Pr_{\phi,\alpha}$ on state sequences. With the assumption, many action sequences induce the same distribution: we obtain $(k_{\text{treat}} + 1)^N$ classes of action sequences, $k_{\text{treat}} = |A_{\text{treat}}|$, where the sequences in each class induce the same distribution. Action sequences that are obtained from one of the admissible plans in the set Π though, contain at most one action choice from A_{treat}. With that restriction, the number of classes therefore further reduces to $N \cdot k_{\text{treat}} + 1$. We will exploit this significant reduction in the solution method described below.

4 Solution method

The standard approach to solving POMDP problems was initiated by Aström [1] and Sondik [11], and is based on transforming the POMDP into an equivalent, fully observable Markov decision process (called the *belief MDP*), over all possible probability distributions on the original state space Ω_X. The belief MDP can be solved using value iteration, a form of dynamic programming [2]. However, the continuous state space of the belief MDP is computationally difficult to handle, and therefore the associated solution algorithms are complicated and limited [8]. Notwithstanding recent algorithmic advances in this field [3; 6], solving POMDP problems of considerable size with this approach seems to be infeasible; the current state of the art allows to solve POMDPs with at most 10 to 15 states. Another disadvantage of dynamic programming is that the decisions are optimised in reverse order. This implies that we cannot exploit prior knowledge of the problem involved (e.g. patient-specific information), and it is difficult to take into account constraints on plan structure, as for instance occur in the specialised POMDP form described above. We therefore propose a new solution method to solve POMDPs, tailored to the specialised form described above. Due to space limitations, we restrict ourselves to giving a sketch of the proposed method.

Basically, our method estimates expectations of the utility function u under a given decision-theoretic plan $\pi \in \Pi$ by simulating the stochastic process on X under plan π. These Monte Carlo estimates are then compared to establish the optimal plan. With this approach, we can easily exploit prior knowledge of the problem case, as each simulation starts from the initial decision moment; this is especially useful when many potential state sequences are ruled out by the initial state. Constraints on plan structure are taken into account by selecting plans from the admissible set Π only. Furthermore, we can take advantage of the fact that the distribution on state sequences is fixed by the choice and timing of treatment. Let $\sigma_1, \ldots, \sigma_n$ be independent and identically distributed samples from $\mathcal{S}$, where treatment action $a \in A_{\text{treat}}$ was selected at time point $t < N$ in the simulations. Since the transition probabilities are equal for all test actions and the skip action, we can use these samples to estimate expectations of the function u for all action sequences that select treatment a at moment t, regardless of their prior testing policy. So, the simulation effort is strongly reduced as we evaluate a large variety of action sequences from a single collection of samples.

The space Π of admissible plans will generally be too large to enumerate. We therefore perform a local search through Π, stepwise refining the plan under consideration. The search process proceeds as follows. Let ϕ represent given beliefs on the initial state, and let α be the action sequence where action $a \in A_{\text{treat}}$ is selected at time point $t < N$, and *skip* is selected at all other times. Note that α also represents a (rather unsophisticated) plan $\pi \in \Pi$: 'perform action a at time point t without prior testing'. Now, let $\Pr_{\phi,\alpha}$ as before be the distribution on $\mathcal{S}$ induced by ϕ and α, and let

S be a collection of independent and identically distributed samples from $\mathcal{S}$ drawn using $\Pr_{\phi,\alpha}$. If $\hat{u}(\sigma,\alpha)$ denotes the life expectancy associated with state and action sequences σ and α, then

$$\bar{u}_{\phi,\alpha}(S) \quad = \quad \frac{1}{|S|}\sum_{\sigma\in S}\hat{u}(\sigma,\alpha) \qquad (4)$$

is an Monte Carlo estimate of life expectancy under plan π. To obtain more sophisticated plans, we now try to find *indicators* of variation in $\bar{u}$. We say that the set $Y \subseteq X$ is such an indicator at time point $t' < t$, if there exists configurations c'_Y and c''_Y of Y such that difference between $\bar{u}_{\phi,\alpha}(S')$ and $\bar{u}_{\phi,\alpha}(S'')$ is statistically significant, where S', S'' are the subcollections of state sequences matching c'_Y and c''_Y at time point t', respectively. We restrict the search process to indicators Y that are *observable*, i.e. $Y = o(a')$ for some action $a' \in A_{\text{test}}$. Furthermore, the difference between estimated life expectancies must remain significant when adjusted for performing test action a' at time point t'. The plan π is now refined by adding the test action corresponding to the indicator that induces the most significant difference in life-expectancy estimates. Subsequently, the treatment action and its timing are re-considered for each of the possible observations that may follow a'; new simulations may be needed to obtain the necessary samples here. After possible adjustment of the treatment choice under each of the observations, the process is repeated; policy refinement is halted when no further improvements can be found.

We note that Monte Carlo estimates converge to correct expected values in the limit of taking an infinite number of samples. In practice, however, a finite, and often small, number of samples is sufficient. Furthermore, the number of samples corresponding to particular events is balanced with the likelihood of these events to occur. In our application of the technique, this means that highly improbable state developments are considered only after taking a large number of samples. At the start of the policy-refinement process, improvements to the policy will be based on developments that are either very likely to occur or induce large differences in life expectancy. As the refinement process proceeds and more samples are obtained, improvements may also be based on rare developments that induce small differences.

5 Discussion and future work

POMDPs provide a powerful modelling framework for decision-theoretic planning, with promising applications to multi-stage clinical decision problems. The generality of the standard POMDP model, however, limits practical application of the framework due to the computational complexity of associated solution methods. To alleviate this obstacle, we have proposed a specialised POMDP form and algorithm to support a frequently encountered type of clinical management problem. The specialised form assumes several restrictions on the effects of actions on state development, and on the structure of admissible solutions. These restrictions jointly reduce the number of action-sequence classes that induce a different probability distribution on state sequences. Our algorithm exploits this property by reducing the simulation effort in Monte-Carlo evaluation of decision-theoretic plans: each sample of the stochastic process is used to evaluate a large number of action sequences.

We are currently implementing our algorithm, and plan to evaluate its performance on the VSD model in the near future. Further research is required to investigate extensions to the basic model form proposed here. For instance, more elaborate loss and utility functions that incorporate quality of life and costs of treatment, are needed to provide a more realistic account of the tradeoffs in real-world clinical decisions. Furthermore, allowing a larger number of control moments is needed to support a wider range of management problems. To prevent a combinatorial explosion in the solution space, this extension should be coped to a fine-grained classification of action types and associated restrictions on admissible treatment plans.

Acknowledgements

The investigations were (partly) supported by the Netherlands Computer Science Research Foundation with financial support from the Netherlands Organisation for Scientific Research (NWO). The author wishes to thank Jaap Ottenkamp for his support in constructing the VSD model.

References

[1] K.J. Aström. Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.*, 10:174–205, 1965.

[2] R.E. Bellman. *Dynamic Programming.* Princeton University Press, 1957.

[3] A.R. Cassandra, M.L. Littman, and N.L. Zhang. Incremental pruning: a simple, fast, exact method for partially observable Markov decision processes. In *Proc. 13th Conf. Uncertainty in Artificial Intelligence (UAI–97)*, pp. 54–61, 1997.

[4] A.W. Drake. *Observation of a Markov Process through a Noisy Channel.* Ph.D. thesis, MIT, 1962.

[5] E. Fernández-Gaucherand and S.I. Marcus. Risk-sensitive optimal control of hidden Markov models: Structural results. *IEEE Trans. Automatic Control*, 42:1418–1422, 1997.

[6] M. Hauskrecht. Incremental methods for computing bounds in partially observable Markov decision processes. In *Proc. 14th Nat. Conf. Artif. Intell. (AAAI–97)*, 1997.

[7] M. Hauskrecht and H. Fraser. Planning medical therapy using partially observable Markov decision processes. In *Proc. 9th Int. WS Principles of Diagnosis (DX–98)*, pp. 182–189, 1998.

[8] W.S. Lovejoy. A survey of algorithmic methods for partially observed Markov decision processes. *Ann. Oper. Res.*, 28:47–66, 1991.

[9] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of Markov decision processes. *Math. Oper. Res.*, 12(3):441–450, 1987.

[10] N.B. Peek. Explicit temporal models for decision-theoretic planning of clinical management. *Artif. Intell. Med.*, 15(2):135–154, 1999.

[11] E.J. Sondik. *The Optimal Control of Partially Observable Markov Processes*. Ph.D. thesis, Stanford University, 1971.

Robust Outcome Prediction for Intensive-Care Patients

Marco Ramoni
Knowledge Media Institute
The Open University
United Kingdom

Paola Sebastiani
Department of Statistics
The Open University
United Kingdom

Richard Dybowski
Intensive Care Group
King's College London
United Kingdom

Abstract

Missing data are a major plague of medical databases in general, and of Intensive Care Units (ICUs) databases in particular. The time pressure of work in an ICU pushes the physicians to omit randomly or selectively record data. These different omission strategies give rise to different patterns of missing data and the recommended approach of completing the database using median imputation and fitting a logistic regression model can lead to significant biases. This paper applies a new classification method, called robust Bayes classifier, that does not rely on any particular assumption about the pattern of missing data and compares it to the traditional median imputation approach using a database of 327 ICU patients.

1 Introduction

The primary role of intensive care units (ICUs) is to monitor and stabilize the vital functions of patients with life-threatening conditions. In order to aid ICU nurses and intensivists with this work, *scoring systems* have been developed to express the overall state of an ICU patient as a numerical value. ICU data sets often have missing values. One suggestion as to why a patient attribute remains unrecorded is that an intensivist assumes the variable to be clinically normal on the basis of some other observation and, therefore, not worthy of confirmation. Although this clinical-normality assumption has been criticized [2], the mortality rate is higher in those patients with completed records. Since abnormal physiological values are associated with increased risk, it has been argued that this supports the clinical-normality assumption. In addition to this, we suspect that there are random omissions due to the pressure of work within an ICU; thus, it may be the case that the incompleteness of an ICU data set is due to a mixture of different missing-data mechanisms. This situation motivates the investigation presented in this paper. We compare a logistic regression model derived from imputed missing data with a Bayesian classifier using a robust Bayesian estimator [10] to handle missing data. The main character of this robust Bayesian classifier is its ability to learn and predict on the basis of incomplete data with no assumption about the missing data mechanism.

2 Prognostic Models

In this section we describe two prognostic models: logistic regression and the Naive Bayesian Classifier.

2.1 Logistic Regression Models

APACHE II [4] is a subjective linear combination based on demographic and physiological attributes, which increases as the state of a patient declines. In spite of its subjectivity, posterior probabilities of a defined outcome have been estimated by having APACHE II as a logistic-regression covariate [3]. In 1985 [7], APACHE II was replaced with the logistic regression model, in which the probability of a patient outcome is modeled as a *logit* function of the attribute values via the function

$$p(outcome|\mathbf{x}) = (1 + \exp[-(w_0 + \sum_{i=1}^{m} w_i x_i)])^{-1} \quad (1)$$

The variable *outcome* is binary, corresponding to the two states alive or not while in ICU, the x_i are attributes that, in the model in Equation 1, are not supposed to interact, and the w_i values are parameters that can be estimated from available data, using Maximum Likelihood estimators. Once the parameters w_i are estimated from a data set of cases, the model in Equation 1 can be used for prediction of a patient outcome, by selecting the outcome with the largest probability, or for defining a number of objective scoring systems, which have proved to perform better than those obtained subjectively [1].

2.2 The Naive Bayes Classifier

Outcome prediction can be transformed into a classification task by regarding the regression covariates as attributes of two alternative classes representing the patient outcome. In this section, we will describe

the application of a Naive Bayes Classifier (NBC) [6; 9] to a ICU database. A NBC is a supervised classification model that assumes the conditional independence of the attributes given the class. We describe the NBC in the context of two classes, although the NBC can be used more generally, when the number of classes is greater than two.

A NBC is defined by the marginal probabilities $\{p_1, p_2\}$ of the two classes and by the conditional probabilities $\{p_{ijk}\}$ of each attribute value x_{ik} given each class c_j. These probabilities can be easily estimated from the data as relative frequencies or adjusted relative frequencies to account for prior information, when the attribute are discrete variables. As the logistic regression model in Equation 1, the NBC can be used to evaluate the probability of a class, given a set of attribute values $e = \{x_{1k}, ..., x_{mk}\}$ as

$$p(c_1|e_k) = \frac{\prod_{i=1}^{m} p_{i1k} p_1}{\prod_{i=1}^{m} (p_{i1k} p_1 + p_{i2k} p_2)}. \qquad (2)$$

This probability is then used for predicting the outcome of a patient on the basis of his/her attribute values or to define some scoring system, as discussed in the previous section.

3 Missing Data

When some entries in the data set are reported as unknown, the estimation of the parameters w_i in the logistic regression model 1 and of the probabilities $\{p_j\}$ and $\{p_{ijk}\}$ in the NBC can be done by using imputation [8]. Imputation essentially consists of replacing the unknown entries by some value generated from an imputation model that depends on the assumption made about the missing data mechanism. Here, we follow the classification introduced by Rubin [12]:

- Data are said to be *missing completely at random* if the probability that an entry is missing in the data set is independent of the other values, observed or not;
- Data are said to be *missing at random* if the probability that an entry is missing in the data set is a function of the values observed in the data set;
- Data are said to be informatively missing if the probability that an entry is missing is a function of the values observed or not in the data set.

In the context of ICU data, the missing entries in the data set are missing completely at random when they are caused by random omissions, as for instance due to work pressure. Data that are omitted due to the assumption of clinical normality can be described as being missing at random, because the intensivist assumes the omitted variables to be clinically normal on the basis of some other observation and, therefore, not worth confirmation. This situation is different from deliberately omitting values that were measured. This last case would yield data that are informatively missing.

The assumption about the missing data mechanism affects the way that either the logistic regression model or the NBC are induced from the available data. When data are missing completely at random then the data available are still a "representative sample", on the whole. When data are missing at random, then the data available are a representative sample when taken by groups. In both cases, the available data are sufficient to fill in — either deterministically or stochastically — the missing entries. When neither of these two assumptions hold, their enforcement can introduce severe bias, and a correct model building relies on the knowledge of the process responsible for the missing data. Sebastiani and Ramoni [13] provide examples of the bias due to an indiscriminate use of imputation.

Clearly, the solution is to use the correct imputation model, but this is not always possible because of lack of information about the process that caused missing data. In the next section, we describe a method for robust classification that does not require any specific model for the missing data mechanism.

4 Robust Classification

The robust Bayesian estimator introduced by Ramoni and Sebastiani [10] is a novel approach that allows to estimate the probabilities $\{p_{ijk}, p_j\}$ specifying the NBC without making any assumption about the missing data mechanism. This feature seems to be the appropriate solution to the complexity of missing data mechanisms involved in ICU databases. This estimator is based on a new view of incomplete data: with no information on the pattern of missing data, an incomplete data set can only constrain the set of estimates that can be induced from the database. Hence, the robust Bayesian estimator returns probability estimates that are robust with respect to the missing data mechanism by providing probability intervals that contain the estimates learned from all possible completions of the incomplete database. The calculation of these interval estimates is done very efficiently by computing virtual frequencies that correspond to extreme completions of the incomplete data. Compared to imputation, the robust Bayesian estimator does not rely on a single model for the missing data, but provides sets of estimates consistent with all possible missing data mechanisms from which the incomplete data at hand could have been obtained.

However, in order to use the estimates computed by the robust Bayesian estimator to produce a robust prognostic model, we need to find a solution to the following problems:

1. the evaluation of the posterior probability in Equation 2 requires the probabilities $\{p_{ijk}, p_j\}$ to be point valued;
2. the use of intervals prevents the use of the standard criterion of selecting the class with the highest posterior probability, because the posterior

probabilities are intervals rather than single values.

Ramoni and Sebastiani [11] describe an exact algorithm for extending Equation 2 to probability intervals. The algorithm maintains the same computational complexity needed to evaluate Equation 2 and returns, for each conditional probability $p(c_j|e)$, probability intervals $[p_{min}(c_j|e); p_{max}(c_j|e)]$ that contain all the values we would obtain from the possible completions of the data.

We now need a method for ranking probability intervals, so that the prediction can be done by choosing the class associated with the highest ranked interval. The stochastic dominance criterion, proposed by Kyburg [5], predicts the class c_j if and only if the minimum posterior probability of this class is higher than the maximum posterior probability of the other classes. Stochastic dominance is the safest and most conservative criterion, as it is independent of the missing data mechanism. However, this criterion is unable to classify cases when intervals are overlapping and we therefore have to resort to a *weak dominance* criterion. Weak dominance summarizes each probability interval into a point, called a *robust classification score*. The score is computed by assuming a uniform distribution over the missing data, and the prediction is done by selecting the class associated with the highest score. More formally, by letting q denote the number of classes, we define the *robust classification score* of $c_j|e$ as

$$s_u(c_j|e) = \frac{p_{min}(c_j|e)(q-1)}{q} + \frac{p_{max}(c_j|e)}{q}$$

When there are two classes, as in the present situation, the score $s_u(c_j|e)$ is the interval mid-point. We note that, in using a uniform distribution over the interval values, we are not assuming that data are missing completely at random. The latter condition would require to weight each maximum probability $p_{max}(c_j|e)$ by the probability p_j that an unobserved entries of C is c_j.

We define a NBC that is induced from incomplete data using the robust Bayesian estimator, and that is used for classification using either stochastic dominance or weak dominance, the robust Bayes classifier. The program RoC[1], implements the robust Bayes classifier.

5 Experimental Evaluation

This section reports an experimental comparison on a ICU database between a logistic regression model and the robust Bayes classifier. We first describe the data set and the procedure used compare the two models.

[1]available at `http://kmi.open.ac.uk/projects/bkd`

5.1 Material and Methods

The 327 patients comprising the data set were present in the adult ICU at St Thomas' Hospital, London, from January 1997 to July 1997. The 11 variables in the data set are listed in Table 1, and the values are those recorded during the first 24-hours of each patient's stay in ICU. The data set is incomplete, of the 11×327 cells of the data set, 75 (2%) are empty, resulting in 67 (20%) incomplete rows.

Contrary to the robust Bayes classifier that does not need any assumption about the missing data mechanism, the estimation of the parameters w_i of the logistic regression model relies on some explicit model for the missing data to allow for imputation. We imputed the missing entries in the data set, under the assumption that data were missing completely at random. Hence, the missing entries of each covariate were replaced by a reference value computed from the marginal distribution of the covariate itself. Since the covariates have skewed distributions, we replaced the missing entries by the observed median of each variable, that is less sensitive to outliers.

The comparison of predictive accuracy of the two models was carried out by running ten replicates of a 5-fold cross validation experiment. In each replicate, we divided the data set in 5 mutually exclusive data sets $\mathcal{D}_1, ..., \mathcal{D}_5$ of approximately the same size. For each data set $\mathcal{D}_i$, we estimated both the logistic regression model and the robust classifier on the data set $\mathcal{D}$ in which we removed the cases in $\mathcal{D}_i$ and we then used the two models to predict the outcome of patients in $\mathcal{D}_i$.

Estimation Each logistic regression model was estimated using the S-Plus `glm` function with the argument `family=binomial`. In each case, we fitted additive logistic regression models without employing interaction terms. Each robust Bayes classifier was estimated using the program RoC that implements the robust classification described in Section 4. Continuous variables were discretized in four equally spaced intervals of the logarithmic transformation of the observed values.

Prediction For this study, a patient is classified as not surviving in hospital if his/her posterior probability for death while in hospital is greater than 0.5 according to the logistic regression model. On the other hand, the robust Bayes classifier under the strong dominance criterion classifies a patient as not surviving if the minimum probability of not surviving is larger than the maximum probability of surviving. The robust Bayes classifier under the weak dominance criterion predicts the patient outcome as that one corresponding to the probability interval with largest mid-point.

As each data set D_i contains the observed outcome, we evaluate the performance of the two models

Table 1: The attributes of interest

Variable name	Data type	Code
Age (years)	Continuous	—
Artificial ventilation required	Nominal	"1" = true; "2" = false
Type of inotrope support	Ordinal	"0" = no intotropes; "1" = dopamine; "2" = adrenaline only; "3" = adrenaline plus other inotrope(s)
Serum bilirubin (mmol/l)	Continuous	—
Acute renal failure	Nominal	"1" = true; "2" = false
24-h urine volume	Ordinal	"0" = (0 - 50ml); "1" = (51 - 300ml) ; "2" = (> 300ml)
Surgical category	Nominal	"1" = elective (mostly cardiothoracic); "2" = emergency (medical patients); "3"= emergency (general surgery)
Creatinine	Continuous	—
Left ventricular intercept	Continuous	—
Glasgow coma score	Ordinal	1,2,...,15
Alive whilst in hospital	Nominal	"1" = true; "2" = false

by comparing their predictive *accuracy* and *coverage*. The predictive accuracy measures the predictive quality as the average number of cases that were correctly classified in the test sets. The coverage is the ratio between the number of cases classified and the total number of cases in the data set. Hence, the coverage of the logistic regression model is 100%, as well as the coverage of the robust Bayes classifier that uses the robust classification score. The coverage of the robust Bayes classifier that uses the stochastic dominance criterion is the ratio between the number of cases that were classified and the size of the data set. Since the cross validation procedure was repeated 10 times, the classification accuracy and coverage are computed as average values of the resulting 10 measures of accuracy and coverage, and we also provide 95% confidence limits.

5.2 Results

The average classification accuracy of the logistic regression model was 80.25% ±2.15. The classification accuracy of the robust Bayes classifier that uses the stochastic dominance criterion increases to 85.4% ±2.05. The price of such increased accuracy is a decreased coverage of 87.76% ±2.06. Using the weak dominance criterion, we increased the coverage of the robust Bayes classifier to 100% by reducing the accuracy to 80.70% ±2.15. Hence, compared to logistic regression, the gain of accuracy is 0.5%.

6 Conclusions

A conservative approach, with no commitment to a particular missing data mechanism, improves the predictive accuracy in our example data set but leaves unclassified a quota of the cases. When we increase the coverage by adopting weaker criteria, the accuracy reduces to a level comparable to the accuracy achieved by logistic regression with median imputation. These findings suggest that, in practical applications, a conservative approach can be taken in order to increase the accuracy of the predictions. The unclassified cases can be left for more careful consideration to a human expert, possibly aided by the predictions obtained under weaker criteria.

Furthermore, the fact that, even under stochastic dominance, the accuracy is limited to 85.4% ±2.05 questions the ability of the models considered to represent the real dependence of the outcome variable on the 10 attributes recorded in the data set. However, building improved logistic regression models from the incomplete data can be seriously biased by the imputation method adopted. The robust Bayes classifier can be improved by selecting relevant attributes on the basis of their predictive relevance, without making assumptions on the missing data mechanism. Preliminary results seem to suggest that a careful selection of attributes having a significant predictive relevance can further increase the accuracy of the robust Bayes classifier.

Acknowledgements

This research was supported by the ESPRIT programme of the Commission of the European Community under contract EP29105 and by equipment grants from Apple Computers and Sun Microsystems.

References

[1] X. Castella, A. Artigas, J. Bion, A. Kari, and The European/North American Severity Study Group. A comparison of severity of illness scoring systems for intensive care unit patients: Re-

sults of a multicenter, multinational study. *Critical Care Medicine*, 23:1327–1332, 1995.

[2] H.R. Champion and W.J. Sacco. Measurement of patient illness severity. *Critical Care Medicine*, 10:552–553, 1982.

[3] R.W.S. Chang, S. Jacobs, and B. Lee. Use of APACHE II severity of disease classification to identify intensive-care-unit patients who would not benefit from total parenteral nutrition. *Lancet*, 1986i:1483–1487, 1986.

[4] W.A. Knaus, E.A. Draper, D.P. Wagner, and J.E. Zimmerman. APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13:818–829, 1985.

[5] H.E. Kyburg. Rational belief. *Behavioral and Brain Sciences*, 6:231–273, 1983.

[6] P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In *Proceedings of the National Conference on Artificial Intelligence*, pages 223–228, San Mateo, CA, 1992. Morgan Kaufman.

[7] S. Lemeshow, D.Teres, H. Pastides, J.S. Avrunin, and J.S. Steingrub. A method for predicting survival and mortality of ICU patients using objectively derived weights. *Critical Care Medicine*, 13:519–525, 1985.

[8] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

[9] S.G. Pauker, G.A. Gorry, J.P. Kassirer, and W.B. Shwartz. Toward the simulation of clinical cognition: Taking a present illness by computer. *The American Journal of Medicine*, 60:981–995, 1976.

[10] M. Ramoni and P. Sebastiani. Robust learning with missing data. Technical Report KMi-TR-28, KMI, The Open University, 1996. Available at http://kmi.open.ac.uk/techreports/KMi-TR-28.

[11] M. Ramoni and P. Sebastiani. Robust Bayes classifiers. Technical Report KMi-TR-82, KMI, The Open University, 1999. Available at http://kmi.open.ac.uk/techreports/KMi-TR-82.

[12] D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[13] P. Sebastiani and M. Ramoni. Model folding for data subject to nonresponse. In *Proceedings of Artificial Intelligence and Statistics 1999*, pages 287–292. Morgan Kaufman, San Mateo CA, 1999.

Prognoses for Multiparametric Time Courses of the Kidney Function

Rainer Schmidt [a], Bernhard Pollwein [b], Lothar Gierl [a]

a) Institute for Medical Informatics and Biometry, University of Rostock, Rembrandtstr. 16/17, 18055 Rostock, Germany, e-mail: rainer.schmidt@medizin.uni-rostock.de

b) Department of Anaesthesiology of the Ludwig-Maximilians University, Munich, Germany

Abstract

In this paper, we describe an approach to utilize Case-Based Reasoning (CBR) methods for trend prognoses for medical problems. Since using conventional methods for reasoning over time does not fit for course predictions without medical knowledge of typical course pattern, we have developed abstraction methods suitable for integration into our Case-Based Reasoning system ICONS. These methods combine medical experience with prognoses of multiparametric courses. We have chosen the monitoring of the kidney function in an Intensive Care Unit (ICU) setting as an example for diagnostic problems. On the ICU, the monitoring system NIMON provides a daily report based on current measured and calculated kidney function parameters. We subsequently generate course-characteristic trend descriptions of the renal function over the course of time. Using Case-Based Reasoning retrieval methods, we search in the case base for courses similar to the current trend descriptions. Finally, we present the current course together with similar courses as comparisons and as possible prognoses to the user. We applied CBR methods in a domain which seemed reserved for statistical methods and conventional temporal reasoning.

1. Introduction

Up to 60% of the body mass of an adult person consists of water. The electrolytes dissolved in body water are of great importance for an adequate cell function. The human body tends to balance the fluid and electrolyte situation. But intensive care patients are often no longer able to maintain adequate fluid and electrolyte balances themselves due to impaired organ functions, e.g. renal failure, or medical treatment, e.g. parenteral nutrition of mechanically ventilated patients. The physician therefore needs objective criteria for the monitoring of fluid and electrolyte balances and for choosing therapeutic interventions as necessary.

At our ICU, physicians daily get a printed renal report from the monitoring system NIMON [1] which consists of 13 measured and 33 calculated parameters of those patients where renal function monitoring is applied. For example, the urine osmolality and the plasma osmolality are measured parameters that are used to calculate the osmolar clearance and the osmolar excretion. The interpretation of all reported parameters is quite complex and needs special knowledge of the renal physiology.

The aim of our knowledge based system ICONS is to give an automatic interpretation of the renal state to elicit impairments of the kidney function on time. That means, we need a time course analysis of many parameters without any well-defined standards. At first glance, this seemed to be a field to apply statistical methods. However, our good results of experiments with a Case-Based Reasoning approach and our investigations of the difficulties to handle multiparametric time course problems without a medical domain theory revealed that CBR methods are more applicable in this field. Although much research has been performed in the field of conventional temporal course analyses in the recent years, none of them is suitable for this problem. Allen's theory of time and action [2] is not appropriate for multiparametric course analysis, because time is represented as just another parameter in the relevant predicates and therefore does not give necessary explicit status [3]. As traditional time series techniques [4] with known periodicities work well unless abrupt changes, they do not fit in a domain characterized by possibilities of abrupt changes and a lack of well-known periodicities at all. One ability of RÉSUMÉ [5] is the abstraction of many parameters into one single parameter and to analyse the course of this abstracted parameter. However, the interpretation of the courses requires complete domain knowledge. Haimowitz and Kohane [6] compare many parameters of current courses with well-known standards. In VIE-VENT [7] both ideas are combined: Courses of quantitative measured parameters are abstracted into qualitative course descriptions, which are matched with well-known standards.

However, in the domain of fluid and electrolyte balance, neither a prototypical approach in ICU settings is known nor exists complete knowledge about the kidney function. Especially, knowledge about the behaviour of the various parameters over time is yet incomplete. So we had to design our own method to deal with course analyses of multiple parameters without prototypical courses and without a complete domain theory (Figure 1).

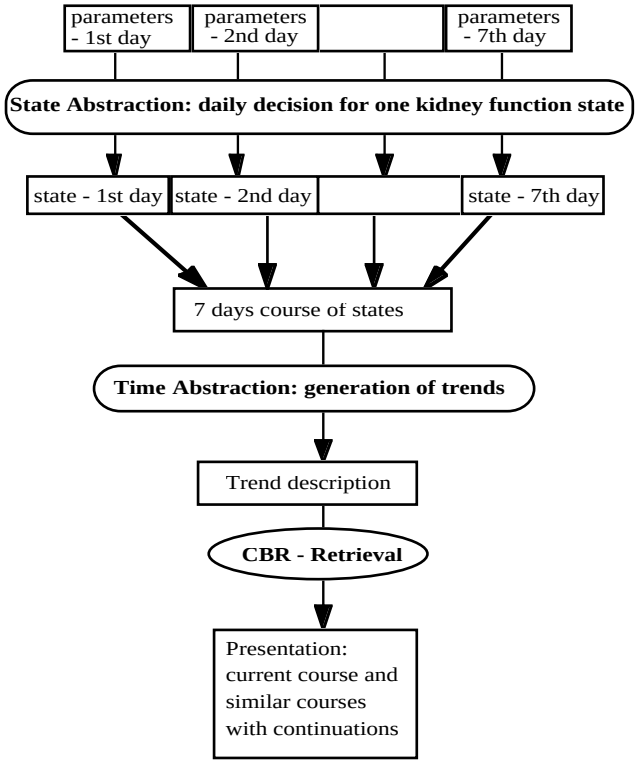

Figure 1 - *Abstractions for Multiparametric Prognoses in ICONS*

2. Methods

2.1. General Model

Our procedure for interpretation of the kidney function corresponds to a general linear model. First, once a day the monitoring system NIMON gets 13 measured parameters from the clinical chemistry and calculates 33 meaningful kidney function parameters. To elicit the relationships among these parameters a three dimensional presentation ability was implemented inside the renal monitoring system NIMON. However, complex relations among all parameters are not visible.

We decided to abstract these parameters. For this data abstraction we use states of the renal function which determine states of increasing severity beginning with a normal renal function and ending with a renal failure. Based on these state definitions, we determine the appropriate state of the kidney function per day. Therefore, we present the possible states to the user sorted according to their probability. The physician has to accept one of them. Based on the transitions of the states of one day to the state of the following day, we generate four different trends. These trends, which are abstractions of time, describe the courses of the states. Then we use Case-Based Reasoning retrieval methods [8, 9, 10, 11] to search for similar courses. We present similar courses together with the current one as comparisons to the user, the course continuations of the similar courses serve as prognoses.

As there may be too many different aspects between both patients, the adaptation of the similar to the current development is not done automatically. ICONS offers only diagnostic and prognostic support, the user has to decide about the relevance of all displayed information. When presenting a comparison of a current course with a similar one, ICONS supplies the user with the ability to access additional renal syndromes and the development of single parameter values during the relevant time period.

2.2. Determination of the Kidney Function State

Based on the kidney function states, characterized by obligatory and optional conditions for selected renal parameters, first we check the obligatory conditions. For each state that satisfies the obligatory conditions we calculate a similarity value concerning the optional conditions. We use a variation of Tversky 's [8] measure of dissimilarity between concepts. If two or more states are under consideration, ICONS presents these states sorted to the similarity values together with information about the satisfied and not satisfied optional conditions (Figure 2).

The user can accept or reject a presented state. When a suggested state has been rejected, ICONS selects another state. The choice depends not only on the computed similarity value, but also on previous decisions of the user and the relation between the states. The states are ordered according to the grade of renal impairment, e.g. it is not necessary to present the state "reduced kidney function", if the user has already accepted the state "sharply reduced kidney function". Finally, we determine the central state of occasionally more than one states the user has accepted. This central state is the closest one towards a kidney failure. Our intention is to find the state indicating the most profound impairment of the kidney function.

2.3. Course-characteristic Trend Descriptions

First, we have fixed five assessment definitions for the transition of the kidney function state of one day to the state of the following day. These assessment definitions are related to the grade of renal impairment:

steady: both states have the same severity value.
increasing: exactly one severity step in the direction towards a normal function.

sharply increasing: at least two severity steps in the direction towards a normal function.
decreasing: exactly one severity step in the direction towards a kidney failure.
sharply decreasing: at least two severity steps in the direction towards a kidney failure.

These assessment definitions are used to determine the state transitions from one qualitative value to another. Neighbouring state transitions with the same assessment are combined into trend pieces. Based on these trend pieces, we generate three trend descriptions. Two trend descriptions especially consider the current state transitions. The first trend description T1 is equivalent to the current trend piece, the second trend description T2 looks recursively back from the current trend piece to the one before and unites them, if they are both of the same direction or one of them has a "steady" assessment. A third trend description T3 characterizes the whole considered course of at most seven days. In addition to the five former assessment definitions we introduced four new ones.

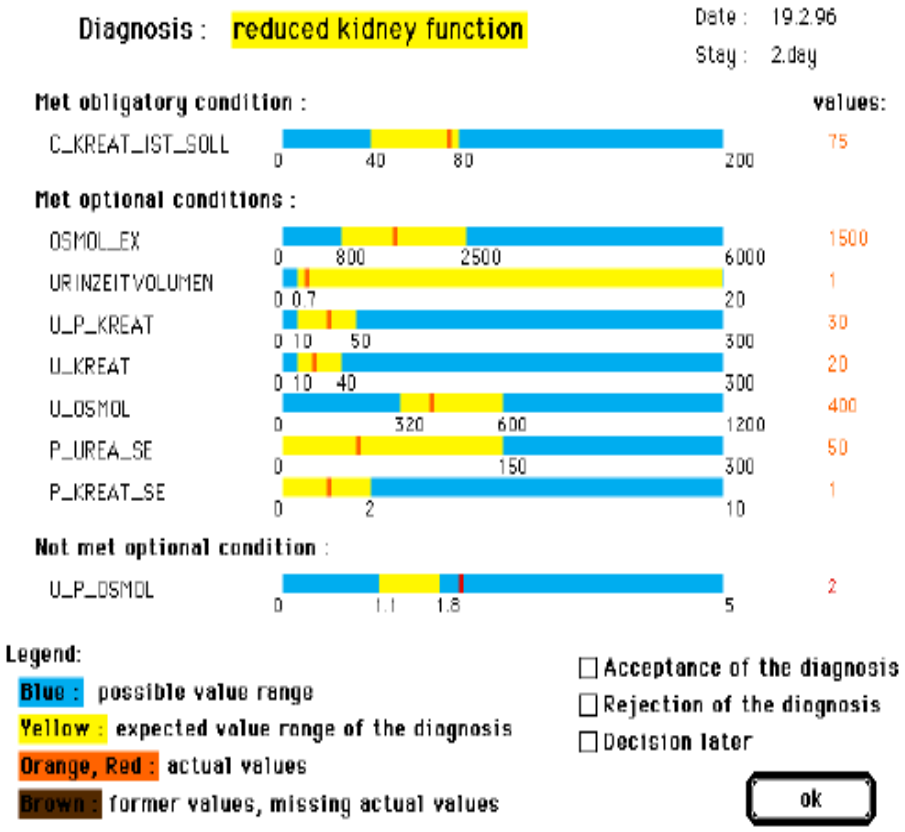

Figure 2 - *Presentation of a current kidney function state estimated as reduced kidney function*

If none of the five former assessments fits the complete considered course, we attempt to fit one of these four definitions in the following order:

alternating: at least two up and two down transitions and all local minima are equal.
oscillating: at least two up and two down transitions.
fluctuating: the distance of the highest to the lowest severity state value is greater than 1.
nearly steady: the distance of the highest to the lowest severity state value equals one.

A fourth trend description T4 assesses the complete considered course with a quantitative value that expresses the average number of state transition values inversely weighted by the distance to the current day.
Looking back from a time point t, these four trend descriptions form a pattern of the immediate course history of the kidney function considering qualitative and quantitative assessments.

2.4. Retrieval

We use the parameters of the four trend descriptions and the current kidney function state to search for similar courses (a retrieval result is depicted in Figure 3). As the

aim is to develop an early warning system, we need a prognosis. For this reason and to avoid a sequential runtime search along the whole cases, we store a course of the previous seven days and a maximal projection of three days for each day a patient spent on the ICU.

As there are many different possible continuations for the same previous course, it is necessary to search for two items: Similar courses and different projections. Therefore, we divided the search space into nine parts corresponding to the possible continuation directions. Each direction forms an own part of the search space. During the retrieval these parts are searched separately and each part may provide at most one similar case. The similar cases of these parts together are presented in the order of their computed similarity values.

Before the main retrieval, we search for a prototype that matches most of the trend descriptions. Below this prototype the main retrieval starts. It consists of two steps for each part. First we search with an activation algorithm concerning qualitative features. Our algorithm differs from the common spreading activation algorithm [9] mainly due to the fact that we do not use a net for the similarity relations. Instead, we have defined explicit activation values for each possible feature value. This is possible, because on this abstraction level there are only ten dimensions with at most six values.

Subsequently, we check the retrieved cases with an adaptability criterion [10], which looks for sufficient similarity, since even the most similar course may differ from the current one significantly. This may happen at the beginning of the use of ICONS, when there are only a few cases known to ICONS, or when the current course is rather exceptional. Because of the lack of medical knowledge about sufficient similarity, we defined a minimal similarity criterion that may be improved after some experience with ICONS.

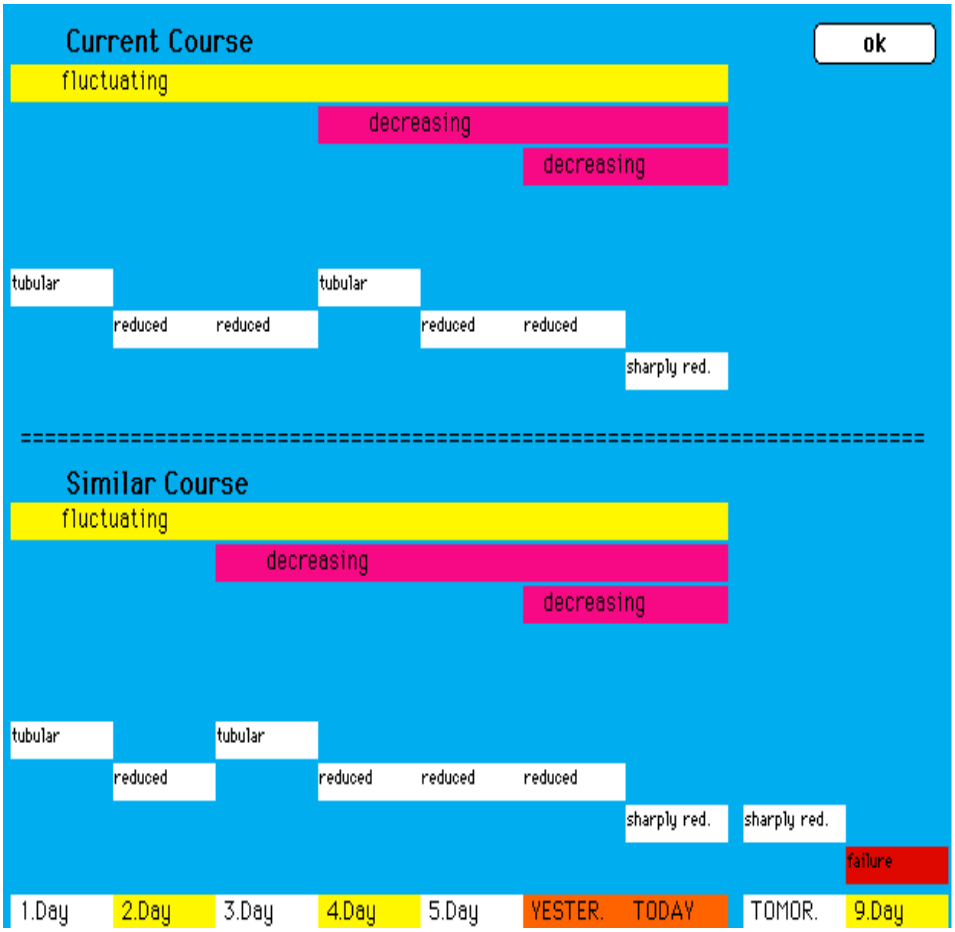

Figure 3 - *Screendump of a comparative presentation of a current and a similar course. In the lower part of each course the (abbreviated) kidney function states are depicted. The upper part of each course shows the deduced trend descriptions.*

If several courses are selected in the same projection part, we use a sequential similarity measure concerning the quantitative features in a second step. This measure is a variation of TSCALE [11] and goes back to Tversky [8].

2.5. Learning a Tree of Prototypes

Prognosis of multiparametric courses of the kidney function for ICU patients is a domain without a medical theory. Moreover, we can not expect such a theory to be formulated in the near future. So we attempt to learn prototypical course pattern. Therefore, knowledge on this domain is stored as a tree of prototypes with three levels and a root node. Except for the root, where all not yet united courses are stored, every level corresponds to one of the trend descriptions T1, T2 or T3. As soon as enough courses that share another trend description are stored at a prototype, we create a new prototype with this trend. At a prototype at level 1, we unite courses that share T1, at level 2, courses that share T1 and T2 and at level 3, courses that share all three trend descriptions. We can do this, because regarding their importance, the three trend descriptions T1, T2 and T3 refer to hierarchically related time periods. T1 is more important than T2 and T3.
We start the retrieval with a search for a prototype that has most of the trend descriptions with the current course in common.

2.6. Evaluation

To verify the knowledge base we selected 100 data sets from the NIMON database. The selection was only partly at random, because we wanted adequate representation of all kidney function states. Two physicians experienced with the kidney function were asked to classify the selected data sets according to the concepts, but without knowing ICONS`s obligatory and optional conditions of the kidney function states. We compared the results of the physicians with ICONS`s classifications of the same data sets. The comparison was mostly satisfactory. For 83 parameter sets the classifications of ICONS corresponded to those of the physicians. In 16 cases ICONS tended more towards the direction of kidney failures. Only once ICONS classified the parameter set as a "reduced kidney function" while the physicians assessed it as a "kidney failure". However, as a result of the evaluation we slightly modified the state definition of the "reduced kidney function".

3. Conclusion

Our aim is to produce an early warning system that helps to avoid kidney failures. ICONS helps the physicians to abstract from the measured and calculated NIMON parameters to a function state. For time periods up to seven days, we describe courses of function states using four trend descriptions as a second abstraction step. At this double abstraction level, ICONS provides the physicians with courses of other patients with similar developments as potential warnings. As no prototypical courses towards a kidney failure are known, we search for cases with similar courses and present them as possible prognoses. We hope to find some prototypical courses by merging similar courses into prototypes. One advantage of combining temporal course analyses with Case-Based Reasoning is the projection. Without medical knowledge about possibilities and probabilities of future developments ICONS shows future developments of patients with similar courses.

References

[1] U. Wenkebach, B. Pollwein, U. Finsterer, " Visualization of large datasets in intensive care", in: Proc Annu Symp Comput Appl Med Care, 1992, pp. 18-22

[2] J.P. Allen, "Towards a general theory of action and time", Artificial Intelligence Vol. 23, 1984, pp. 123-154

[3] E.T. Keravnou E.T., "Modelling Medical Concepts as Time Objects", in: P. Barahona, M. Stefanelli, J. Wyatt (Eds.): Artificial Intelligence in Medicine, AIME‘95, Berlin, 1995, pp. 67-78

[4] S.M. Robeson, D.G. Steyn, "Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations", Atmospheric Environment, Vol. 24 B, No. 2, 1990, pp. 303-12

[5] Y. Shahar, M.A. Musen, "RÉSUMÉ: A Temporal-Abstraction System for Patient Monitoring", Computers and Biomedical Research, Vol. 26, 1993, pp. 255-273

[6] I.J. Haimowitz, I.S. Kohane, "Automated Trend Detection with Alternate Temporal Hypotheses", in: R. Bajcsy (ed.), Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), Morgan Kaufmann, San Mateo, CA, 1993, pp. 146-151

[7] S. Miksch, W. Horn, C. Popow, F. Paky, "Therapy Planning Using qualitative Trend Descriptions", in: P. Barahona, M. Stefanelli, J. Wyatt (Eds.): Artificial Intelligence in Medicine, Springer Berlin, 1995, pp. 197-208

[8] A. Tversky, "Features of Similarity", Psychological Review, Vol. 84 , 1977, pp. 327-352

[9] J.R. Anderson, "A theory of the origins of human knowledge", Artificial Intelligence, Vol. 40, Special Volume on Machine Learning, 1989, pp. 313-351

[10] B. Smyth, M.T. Keane M.T., "Retrieving Adaptable Cases: The Role of Adaptation Knowledge in Case Retrieval" in: First European Workshop on Case-Based Reasoning (EWCBR-93), 1993, pp. 76-81

[11] W.S. DeSarbo, M.D. Johnson, A.K. Manrei, L.A. Manrai, E.A. Edwards, "TSCALE: A new multidemensional scaling procedure based on Tversky's contrast model", Psychometrika, Vol. 57, 1992, pp. 43-69

Impact of Machine Learning to the Diagnosis and Prognosis of First Cerebral Paroxysm

Igor Zelič, Nada Lavrač
Department of Intelligent Systems
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: igor.zelic@ijs.si, nada.lavrac@ijs.si

Peter Najdenov, Zvonka Rener-Primec
Department of Child Neurology
University Medical Center,
Children Hospital,
Vrazov trg 1, 1000 Ljubljana, Slovenia
e-mail: peter.najdenov@guest.arnes.si,
zvonka.rener@mf.uni-lj.si

Abstract

Children and adolescents referred to clinical examinations due to the loss of consciousness, falls or other paroxysmal motor phenomena are frequently incorrectly diagnosed as epileptic. It is extremely important to avoid incorrect diagnosis of epilepsy, not only because of the stigma attached to the label `epileptic' but also due to frequent side-effects of anti-epileptic drugs which should not be given to patients who cannot benefit from their use. An expert such as an experienced pediatric neurologist is able to make correct diagnosis by a detailed history examination. Since expert neurologists are not always available, we tried to model such expert knowledge by the use of machine learning on data collected from the history of 72 patients: 44 patients having first epileptic seizure and 28 patients having syncope. The induced decision tree suggests a stepwise diagnostic process using just few decisive data entries obtained from the history of the event, resulting in 90% accurate diagnosis of either first epileptic seizure or syncope, with 86% specificity and 93% sensitivity. Using such a decision tree model in practice would provide great help and would result in a much more accurate diagnostic procedure than the rather ill-defined procedures used by inexperienced doctors. Moreover, correct diagnosis made on the basis of history data may help to avoid many further unnecessary investigations. When the machine induced expert knowledge is used as a prognostic model, the syncope outcome means a great relief to the patient, since syncope can easily be prevented by simple measures. On the other hand, the case of epileptic fit prognosis requires a cautious approach with some predictive measures and guidelines about everyday life.

Keywords: Machine learning, decision tree induction, cerebral paroxysm, epileptic seizure, syncope

1 Introduction

The term epileptic seizure refers to a sudden change in the electrical activity of the brain, usually accompanied by subjective or objective changes of behavior, motor activity, posture etc. Epilepsy is a condition with recurrent unprovoked seizures or, in some cases, provoked only by everyday stimuli. Non-epileptic seizures are sudden changes in behavior which are not due to an independent sudden change in the electrical activity of the brain, but are due to a sudden change (drop) of blood flow through the brain; these are also named anoxic (without oxygen) seizures. Non-epileptic seizures can be divided into physiological non-central nervous system events, such as syncope, toxic and psychogenic seizures, etc. The most frequent and still too many times misdiagnosed as epileptic fit are syncopes (fainting fits), which usually begin with sudden fall and are also often quite dramatic, associated with convulsions and incontinence, which may misleadingly suggest epileptic seizure.

Everyday clinical experience in outpatient clinic and in hospitals indicate that the most difficult problem is the distinction between anoxic and epileptic seizure, between fainting fits and epileptic fits, when the first event is sudden fall.The diagnosis of cerebral paroxysmal disorders (fits or faints) is based on detailed patient's history. The objective is to elicit a sequence of events as described by a patient and by a witness, circumstances of the event, stimulus, onset, duration of unconsciousness and immediate postictal course. The goal of a rational diagnostic algorithm is to establish the final diagnosis and to offer the patient an appropriate treatment. Only few investigations are necessary, depending on the syncope or epilepsy diagnosis. If the later is the case, more investigations will follow; therefore correct diagnosis, which is based on detailed and reliable patient's history is a very important starting point.

Clinical features in anoxic seizures are usually recorded in the following sequence: subjective feeling of lightheaded, blurred vision or darkness in front of the eyes (but can still hear voices), pale, cold skin, atonia or limpness (sudden fall), stiffening, jerks, disorientation, incontinence. When signs and symptoms of the event are correctly asked and given by detailed history, an expert (like an experienced pediatric neurologist) is able to make a correct diagnosis from the history data alone. In our study we used medical data about patients admitted to the Children Hospital of Ljubljana in 1997-1998 for the diagnosis of the first paroxysmal event. Out of 72 patients, 44 had first epileptic seizure and 28 patients had syncope. Each patient record consisted of values of 23 attributes, including subjective symptoms just before the event (blurred vision or darkens in front of eyes, sudden fall from vertical position, stiffening, limpness, change in skin color, change of position before seizure) and after event sings (headache, vomiting after episode, long sleep or normal activity), number of previous seizures, time to recover, …

2 Data analysis tools

In this study two machine learning data analysis tools were used: a decision tree learner Magnus Assistant and the Naive Bayesian classifier.

2.1 Magnus Assistant

Magnus Assistant [Mladenić 1990] is a descendant of Assistant [Cestnik et al. 1987] and belongs to the ID3 family of systems for top-down induction of decision trees [Quinlan 1986]. The system recursively builds a binary decision tree. The nodes of the tree correspond to attributes, and leaves (terminal nodes) to diagnostic/prognostic classes. In each recursive step of decision tree construction, the `most informative' attribute (an attribute that minimizes the expected number of tests needed for the classification of new cases) is selected and a subtree is built.

The system's distinctive feature is the handling of noisy data using postpruning aimed at increasing the predicted classification accuracy on unseen cases. To do so, the predicted classification accuracy, estimated by the so-called m-estimate of probability [Cestnik 1990] (see also Section 3.2), of each internal node is compared with the expected accuracy of its subtrees to decide whether to prune the subtrees. For the lack of space, another noise handling mechanism called pre-pruning is not described here.

To classify a new case, a path from the root of the tree is selected on the basis of the values of attributes of the new patient to be classified. In this way, for a given patient record, the path leads to a leaf that determines the class: if the leaf is labeled with more than one class, each with the probability of class prediction, then the class with the highest probability is selected for the classification of a new patient.

The entire decision tree reflects the detected regularities in the data, describing the properties that are characteristic for the subsets of examples belonging to subtrees. The ordering of attributes (from the root towards the leaves of the tree) reflects also the importance of attributes for the outcome class in the leaf. The measure of attribute informativity is the selected measure of importance.

2.2 Naive Bayesian classifier

The Bayesian classifier uses the naive Bayesian formula to calculate the probability of each class *Cj* given the values of all attributes of a given instance to be classified [Kononenko 1991a, Kononenko 1993]. For a n-tuple of values *(V1 … Vn)* of the example to be classified and assuming the conditional independence of the attributes given the class, the conditional probability *p(Cj | V1..Vn)* is calculated as follows:

$$p(C_j \mid V_1 \ldots V_n) = p(C_j) \cdot \prod_i \frac{p(C_j \mid V_i)}{p(C_j)}$$

A new instance is classified into the class with the maximal probability. We use the m-estimate [Cestnik 1990] for computing the estimate of conditional probabilities:

$$p(C_j \mid V_i) = \frac{N(C_j \,\&\, V_i) + m \cdot p(C_j)}{N(V_i) + m}$$

where *N(Cond)* stands for the number of examples for which *Cond* is fulfilled, and *m* is a user-defined parameter. The parameter *m* trades-off the contribution of the relative frequency and the prior probability. The default value *m=2* empirically gives good results.

Addition to original Bayesian classifier is use of fuzzy bounds: continuous attributes have to be prediscretized in order to be used by the Bayesian classifier. The task of discretization is the selection of a set of boundary values that split the range of a continuous attribute into a number of intervals which are then considered as discrete ordered values of that attribute. Discretization can be done manually by a domain expert or by applying a discretization algorithm. Fuzzy bounds discretizate the values of the continuous attributes (or, equivalently, the boundary values) as fuzzy values instead of point values [Kononenko1993].

3 Analysis of results

Table 1 gives the results of data analysis. All the results reported are in terms of the number of correct classifications, classification accuracy, informativity, specificity and sensitivity (specificity and sensitivity are computed for syncope being the negative class and epileptic seizure being the positive class).

Sensitivity is defined as the fraction of positive cases that are correctly classified as positive:

$$Sens = \frac{TruePositive}{TruePositive + FalseNegative}$$

and specificity measures the fraction of negative cases correctly classified as negative:

$$Spec = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

Informativity is defined as the ratio of the average information score of the answers of a classifier on a testing set and the entropy of the prior distribution of classes [Kononenko 1991b].

Using the so-called leave one out evaluation method, results are computed from 72 experiments in which 71 patient records were in turn used for training and one record was used for validating the result. For instance, using the naive Bayesian classifier (with parameter m=2 and m=10), 67 out of 72 cases were correctly diagnosed, amounting to an accuracy of 93% - these are the best results achieved.

Despite the fact that best predictions were achieved using the naive Bayesian classifier, the most interesting results from the medical point of view were achieved using the Magnus Assistant decision tree learner. The decision tree induced from the set of 72 patient records using postpruning (m=10) is, according to the expert neurologist, a very appropriate diagnostic model that indeed reflects the medical knowledge that can be used for distinguishing between epileptic seizure and syncope. This decision tree is shown in Figure 1.

	No. correct	Accuracy	Informativity	Specificity	Sensitivity
Assistant, no pruning	63	87.5	0.85	81	86
Assistant, prepruning 2, 80, 5	63	87.5	0.59	79	95
Assistant, postpruning, m=2	63	87.5	0.85	81	93
Assistant, postpruning, m=10	65	90.28	0.83	86	93
Naive Bayes, fuzzy bounds, m=2	65	90.28	0.77	84	95
Naive Bayes, m=2, m=10	67	93.06	0.8	90	95

Table 1: Results of experiments.

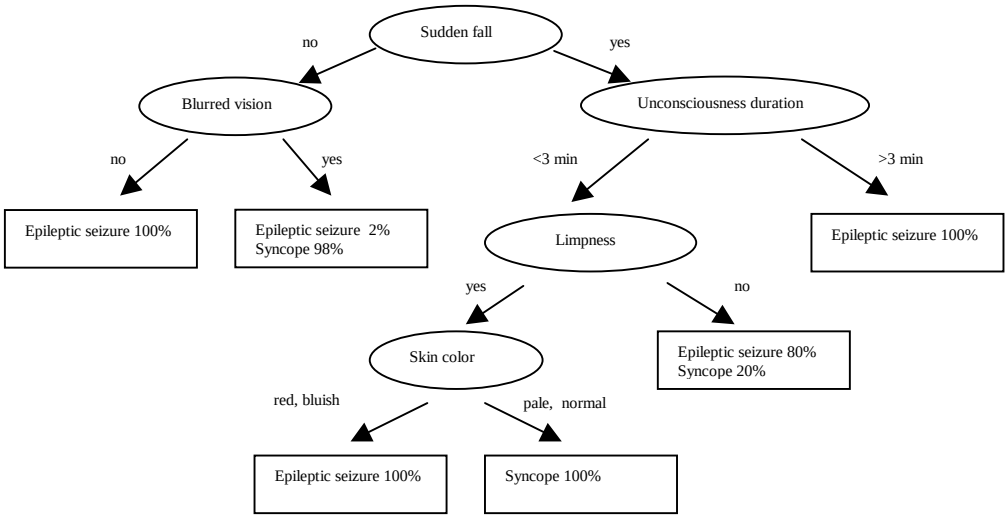

Figure 1: Decision tree induced by Magnus Assistant (with postpruning, m=10)

The most important and usually the most dramatic clinical sign is sudden fall, which is indeed listed at the top of the decision tree and has therefore to be considered as first when distinguishing between the two diagnoses. If there is no fall and the vision is not blurred, the decision tree suggests the diagnosis of epileptic seizure with 100% probability, but if the vision is blurred there is a much larger probability of syncope (98%). When there is sudden fall, the next question asked is the duration of unconsciousness, an item that

has, medically speaking, very high level of specificity. If unconsciousness is longer than 3 minutes, the decision tree suggests epileptic fit with 100% probability; if shorter than 3 minutes, the decision is based on two other attributes: limpness and skin color. The duration of 3 minutes, discovered as a distinguishing value for unconsciousness by the decision tree, is very interesting for an expert neurologist; it has already been observed in practice, but never explicitly recognized as indicative.

4 Discussion and conclusion

In everyday clinical practice in outpatient clinics and hospitals doctors are daily challenged to make as correct diagnosis as possible and perform as few investigation as absolutely necessary because of a huge burden of raising economic costs in medicine as well as because of aiming to decrease the inconvenience for the patient. For many diseases diagnostic algorithms were made by experienced specialists on the basis of many patients they diagnosed in their practice. In our study we evaluated history data from 72 patients, referred either for the first epileptic fit or syncope. It turned out that this data, which seem so obvious to expert neurologists that no one paid much attention to its gathering and analysis, now – by using a machine learning algorithm - turns out to be very appropriate for automatically building a decision making diagnostic algorithm. The induced decision tree can be of great help in diagnosing new cases, especially regarding its prognostic impact: as we stated above it is of outmost importance that a doctor is able, on the basis of history data, with high specificity and sensitivity (as proved by this model) to assure the patient that he is having a benign condition (syncope) that is easily managed by some simple measures and has good prognosis. In the prognosis for such a patient it is extremely important to avoid making a wrong diagnosis of epilepsy, not only because of stigma attached to the label of epileptic but also because of the side effects of anti-epileptic drugs that are frequent and should not be given to patients who cannot benefit from their use.

Another important contribution of the machine learning process using the induced decision tree is a different estimation of some clinical attributes, that are in everyday routine work assessed in some conventional manner. Now, some attributes could be recognized as more informative than previously believed for the diagnosis or prognosis for a patient (e.g., 3 min limit on the duration of unconsciousness). The use of such diagnostic programs can also encourage a clinician to generate new hypotheses and thus aim at the improvement of standard diagnostic and prognostic processes.The presented research is a first step in broader research, which will also include new patient data from the Children Hospital and two general hospitals. Our aim is to collect data for at least 400 patients.

Research will also continue on prognostic model, where we are planning to observe different symptoms of syncope in first year of life (like age, frequency of seizures, duration of seizure, drug effects, etc.) and make a prognostic model of disease development.

References

[Cestnik 1987] B. Cestnik, I. Kononenko, I. Bratko: Assistant 86: A knowledge elicitation tool for sophisticated users. In Machine Learning – Proc. European Working Session on Machine Learning, EWSL-97, Sigma Press, Wilmslow,1987.

[Cestnik 1990] B. Cestnik: Estimating probabilities: A crucial task in machine learning. In Proc. European Conference on Artificial Intelligence, 147-149, 1991.

[Kononenko 1991a] I. Kononenko: Semi-naive Bayesian classifier. In Proc. European Working Session on Machine Learning, EWSL-91, Springer, 206-219, 1991.

[Kononenko 1991b] I. Kononenko, I. Bratko: Information-based evaluation criterion for classifier's performance. Machine Learning, 6(1): 67--80.

[Kononenko 1993] I. Kononenko: Inductive and Bayesian learning in medical diagnosis. Applied Artificial Intelligence 7: 317-337, 1993.

[Mladenić 1990] D. Mladenić: The learning system Magnus Assistant. BSc Thesis. Faculty of Computer and Information Sciences, University of Ljubljana, 1990.

[Michie1994] D. Michie, D.J. Spiegelhalter, C.C. Taylor: Machine Learning, Neural and Statistical Classification. Ellis Horwood, Chichester, 1994.

[Pilih 1997] I.A. Pilih, D. Mladenić, N. Lavrač, T.S. Prevec: Data analysis of patients with severe head injury. In N. Lavrač, E. Keravnou, B. Zupan: Intellignet Data Analysis in Medicine and Pharmacology, Kluwer Academic Press, 131-148, 1997.

[Stephenson 1990] J.B.P. Stephenson: Anoxic seizures or syncopes. In J.B.P. Stephenson: Fits and Faints, Mac Keith Press, Oxford, Blackwell Scientific Publications Ltd., 41-58, 1990.

[Quinlan 1986] J.R. Quinlan: Induction of decision trees. Machine Learning, 1:81-106, 1986.