Author(s): Ramoni Marco, Sebastiani Paola, Dybowski, Richard.
Article Title: Robust outcome prediction for intensive care patients
Year of publication: 2001
Citation: Ramoni M, Sebastiani P, Dybowski R. (2001) "Robust outcome prediction for intensive care patients". Methods of Information in Medicine, 40(1), 39-45.
Link to published version: http://www.schattauer.de/en/magazine/subject-areas/journals-a-z/methods/contents/archive/issue/703/manuscript/128/show.html
DOI: (not stated)
ISSN: 0026-1270

Robust Outcome Prediction for Intensive-Care Patients

Marco Ramoni[1] and **Paola Sebastiani**[2] and **Richard Dybowski**[3]
[1] Knowledge Media Institute, The Open University, United Kingdom
[2] Department of Mathematics, Imperial College, United Kingdom
[3] 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 databases in particular. The time pressure of work in an Intensive Care Unit 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 324 Intensive Care Unit patients.

Keywords: Incomplete Data; Classification; Costs analysis.

Submitted to: Methods of Information in Medicine: Special Issue on Prognostic Models in Medicine – AI and Decision Analytic Approaches.

Corresponding Author: Paola Sebastiani.

Department of Mathematics, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London SW7 2B, United Kingdom. PHONE: +44 ??, FAX: +44 ??, EMAIL: p.sebastiani@ic.ac.uk, URL: http://? .

Contents

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 that is then used to develop a classification rule that classifies a patient as being at risk or not. Such scoring systems typically depend on parameters that are estimated from a database of cases and one feature of these data sets is that, often, they 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 observations 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.

Modeling such a missing data mechanism can be an extremely difficult task and one solution is to resort to a simplifying assumption about the process underlying the missing data. One typical assumption is that data are missing in such a way that the observed data are still a representative sample, on a whole, so that it becomes sensible to replace the missing entries by imputed ones [16]. However, when data are missing in an informative way, such technique can bias the estimate of the parameters of the model used to define the scoring system, thus degrading its classification accuracy.

Based on a completely different approach to modeling incomplete data sets is the theory of robust Bayesian estimation of [14] and its application to classification tasks [13]. The fundamental principle of this theory is that, with no information about the missing data mechanism, an incomplete data set can only constrain the set of estimates that can be induced from all its possible completions and, consequently, classification rules derived from incomplete data sets need to account for this uncertainty. One scoring system that obeys this principle is the Robust Bayes classifier (here on denoted by RBC) introduced in [13]. The uncertainty on the set of estimates, due to the incompleteness of the data sets, implies that the parameters defining a RBC are estimated by probability intervals rather than point-valued probabilities computed under a specific model for the missing data. The second feature of the RBC is its ability to classify cases by reasoning with probability intervals. The interval-based classification is based on a propagation algorithm that computes posterior probability intervals containing all the scoring values that could be obtained from the exact computation of all possible completions of the training set, and one of two possible methods to rank probability intervals. One method is very robust and classify one case only when all assumptions about the missing data mechanism yield the same classification rule. The second method weaken this condition to improve classification coverage but may result in a loss of accuracy. A decision-theoretic criterion then allows one to select the best interval-ranking method by trading

off accuracy and coverage. Extensive evaluations presented in [13] have shown that the RBC is an effective alternative to other scoring systems.

In this paper we compare a scoring system based on a logistic regression model, derived by replacing the missing values with imputed data, with a scoring system based on the RBC. We extend the results in [13] to define a decision-theoretic criterion that takes into account different costs of misclassification and compare the methods in a data set of 324 ICU patients.

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 [5] 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 [8], APACHE II was replaced with the logistic regression model, in which the probability of an ICU patient surviving is modeled as a *logit* function of m covariates $x_1, ..., x_m$ via the function

$$p(s|x_{1k}, ..., x_{mk}) = \frac{\exp[w_0 + \sum_{i=1}^{m} w_i x_{ik})]}{1 + \exp[w_0 + \sum_{i=1}^{m} w_i x_{ik})])}. \quad (1)$$

The event s denotes the ICU patient surviving whilst at the hospital, so that the probability that the patient does not survive, that we will denote by $\bar{s}$, is computed as $1 - p(s|\mathbf{x})$. The x_{ik} are values of the covariates that, in the model in Equation 1, are not supposed to interact, and the values w_i are parameters that can be estimated from available data, using Maximum Likelihood estimators [10]. 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 covariates $x_1, ..., x_m$ as attributes of two alternative classes s and $\bar{s}$ representing the patient outcome. In this section, we focus attention to a Naive Bayes Classifier (NBC) [7, 12] which 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 $\{p(s), 1 - p(s)\}$ of the two patient outcome and by the conditional probabilities $p(x_{ik}|s)$ and $p(x_{ik}|\bar{s})$ of each attribute

value x_{ik} given the two classes s and $\bar{s}$. 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 posterior probability that a patient survives, given a set of attribute values $e_k = \{x_{1k}, ..., x_{mk}\}$ as

$$p(s|e_k) = \frac{\prod_{i=1}^{m} p(x_{ik}|s)p(s)}{\prod_{i=1}^{m}[p(x_{ik}|s)p(s) + p(x_{ik}|\bar{s})(1 - p(s))]}. \quad (2)$$

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 $\{p(s), 1-p(s)\}$ of the two patient outcome as well as the conditional probabilities $p(x_{ik}|s)$ and $p(x_{ik}|\bar{s})$ in the NBC can be done by using imputation [9]. 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 [15] in which 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; and 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, 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. Under the missing completely at random or the missing at random assumption, the data available are still a "representative sample". 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. [17] 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 [14] is a novel approach that allows one to estimate the probabilities $\{p(s), p(x_{ik}|s), p(x_{ik}|\bar{s})\}$ 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 $\{p(s), p(x_{ik}|s), p(x_{ik}|\bar{s})\}$ 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 [13] 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 the probability interval $[\underline{p}(s|e); \overline{p}(s|e)]$ that contains all the values $p(s|e)$ we would obtain from the possible completions of the data. From this interval, one can then derive the probability interval containing all values $1 - p(s|e)$ as $[1 - \overline{p}(s|e); 1 - \underline{p}(s|e)]$.

Ramoni and Sebastiani [13] also proposed two methods for ranking probability intervals, so that the prediction can be done by choosing the patient outcome associated with the highest ranked interval. The first method is based on the strong dominance score which is derived under the stochastic dominance criterion of [6]. The strong dominance score associated with the interval $[\underline{p}(s|e); \overline{p}(s|e)]$ is 1 if and only if the minimum posterior probability $\underline{p}(s|e)$ is higher than the maximum posterior probability $\overline{p}(\bar{s}|e)$ of the other patient outcome, and it is 0 otherwise. If neither of the two conditions is met, the strong dominance score is not defined. In other words, the strong dominance score predicts the outcome of an ICU patient as survival if $\underline{p}(s|e) > \overline{p}(\bar{s}|e)$ and since $\overline{p}(\bar{s}|e) = 1 - \underline{p}(s|e)$,

this reduces to prediction of a patient survival if $\underline{p}(s|e) > 0.5$. If $\overline{p}(s|e) < 0.5$, then the prediction is that the patient will not survive while, if $\underline{p}(s|e) < 0.5 < \overline{p}(s|e)$, the patient outcome cannot be predicted. Strong dominance is a robust 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 weaken it in order to gain classification ability. The second interval ranking method proposed in [13] makes the minimal assumption that all missing data mechanisms are equally possible to define the *complete-admissible score* associated with the interval $[\underline{p}(s|e); \overline{p}(s|e)]$

$$s_u(s|e) = \frac{\underline{p}(s|e) + \overline{p}(s|e)}{2}.$$

From the score $s_u(s|e)$, we derive the score $s_u(\bar{s}|e) = 1 - s_u(s|e)$ associated with the probability interval $[\underline{p}(\bar{s}|e); \overline{p}(\bar{s}|e)]$. This score predicts survival of a patient if $s_u(s|e) > s_u(\bar{s}|e)$ and, hence, if $s_u(s|e) > 0.5$.

Either the strong dominance or the complete-admissible score provide a sensible basis for robust classification. They both have *pros* and *cons*: strong dominance is safe at the price of leaving the outcome of some patients unclassified; the complete-admissible score increases the classification ability with the risk of loosing robustness. We can provide a rule to help one choose the best interval-based classification strategy. The intuition behind this rule is that accuracy is more valuable than coverage and, hence, we would not prefer a method that predicts a patient outcome randomly just because it always makes a prediction. The rationale is that we expect the consequence of a wrong prediction to be worse than the inability to make an automated prediction. This argument can be used formally, to help one choose between the strong dominance or the complete-admissible score by introducing misclassification costs and costs incurred for the inability to classify one case.

We begin by noting that the goodness of a classification system is typically measured via the classification accuracy θ and the coverage γ. The former is the probability of correctly classifying a case while the latter is the probability of classifying one case. Let θ_d denote the accuracy of the RBC with the strong dominance score, say RBC_d, and let γ_d be its coverage. Similarly, let θ_u be the accuracy of the RBC with the complete-admissible score, say RBC_u. Suppose that the cost incurred for not being able to classify a patient is a quantity C_i, while the cost for a wrong classification is C_w. Since the former event occurs with probability $1 - \gamma_d$ and the latter occurs with probability $(1 - \theta_d)\gamma_d$, the expected cost incurred on using the RBC_d is

$$C(\text{RBC}_d) = C_w(1 - \theta_d)\gamma_d + C_i(1 - \gamma_d)$$

if correct classification has no associated costs. On the other hand, the expected cost incurred on using the RBC_u, achieving 100% coverage with accuracy θ_u, is

$$C(\text{RBC}_u) = C_w(1 - \theta_u).$$

If we decide to use the system with minimum expected cost, the RBC_d is to be preferred to the RBC_u when $C(\text{RBC}_d) \leq C(\text{RBC}_u)$ that is true if and only if

$$\theta_u - \theta_d\gamma_d \leq (1-\gamma_d)(1-C_i/C_w).$$

For example, if $C_i = C_w$, the best decision is to choose the RBC_d whenever $\theta_u \geq \theta_d\gamma_d$. In practical applications, the quantities θ_d, θ_u and γ_d can be estimated from the data available by running some cross validation experiment [4].

This principled way to choose the classification system with minimum expected cost can also be used to help one compare other methods. For example, in Section 2, we described logistic regression as the current model to define a scoring system used for predicting whether an ICU patient is at risk of death or not. Suppose the quantity θ_l is the accuracy of logistic regression based classification, with expected cost $(1-\theta_l)C_w$. The comparison between the accuracies θ_l and θ_u is cost-independent, as we compare $C(\text{RBC}_u) = C_w(1-\theta_u)$ with $C_w(1-\theta_l)$ and the minimum expected cost is achieved by the system having the highest accuracy. If we now compare the expected costs of the RBC_d and logistic regression, we have that the latter is to be preferred whenever $\theta_l - \theta_d\gamma_d \geq (1-C_i/C_w)$.

This principled way to compare classification systems is based on the assumption that the cost incurred in classifying an ICU patient as at risk of dying when he is not is equal to the cost incurred in classifying an ICU patient as not being at risk of dying when he is. In real life, the two costs are different and we can describe them in the cost matrix below

Patient true outcome	Patient predicted outcome	
	Survives	Not Survive
Survive	0	$C_{s\bar{s}}$
Not Survive	$C_{\bar{s}s}$	0

In this cost matrix, the quantity $C_{s\bar{s}}$ is the cost incurred in predicting death of a patient who survives while $C_{\bar{s}s}$ is the cost incurred in predicting the survival of a patient who then dies. Let θ_{ss} denote the probability of predicting the survival of a patient who indeed does and let $\theta_{\bar{s}\bar{s}}$ be the probability of predicting the death of a patient who unfortunately dies. The overall accuracy θ of a system can be break down into

$$\theta = \theta_{ss}p(s) + \theta_{\bar{s}\bar{s}}(1-p(s)) \tag{3}$$

where $p(s)$ denotes the prior probability that a patient survives. In words, Equation 3 expresses the accuracy θ as the weighted sum of the probability of predicting the event s, given that s will occur, and of the probability of predicting the event $\bar{s}$, given that $\bar{s}$ will occur. The quantity $1-\theta_{ss}$ and $1-\theta_{\bar{s}\bar{s}}$ can be used to compute the overall expected costs incurred in using the RBC_d as

$$C(\text{RBC}_d) = \{C_{s\bar{s}}(1-\theta_{d,ss})p(s) + C_{\bar{s}s}(1-\theta_{d,\bar{s}\bar{s}})(1-p(s))\}\gamma_d + (1-\gamma_d)C_i$$

while, for example, the overall expected cost incurred in using the RBC_u is given by

$$C(\text{RBC}_u) = C_{s\bar{s}}(1-\theta_{u,ss})p(s) + C_{\bar{s}s}(1-\theta_{u,\bar{s}\bar{s}})(1-p(s))$$

and the overall expected cost incurred in using logistic regression is given by

$$C(\text{Logistic}) = C_{s\bar{s}}(1-\theta_{l,ss})p(s) + C_{\bar{s}s}(1-\theta_{l,\bar{s}\bar{s}})(1-p(s))$$

The choice of the best classification system can then be let depend on a cost analysis. The quantities θ_{ss} and $\theta_{\bar{s}\bar{s}}$ can be estimated from the *sensitivity* and *specificity* of the classification system, that are typically derived from the confusion matrix below

Patient true outcome	Patient predicted outcome		
	Survives	Not Survive	Total
Survive	$n(s,s)$	$n(s,\bar{s})$	$n(s)$
Not Survive	$n(\bar{s},s)$	$n(\bar{s},\bar{s})$	$n(\bar{s})$

The value $n(s,s)$ and $n(\bar{s},\bar{s})$ represent the frequencies of cases correctly classified in a test set (or a cross-validation experiment) with a global number of $n = n(s)+n(\bar{s})$ cases, while $n(s,\bar{s})$ and $n(\bar{s},s)$ are the frequencies of wrong classifications divided according to the type of classification error made. The ratio $n(s,s)/n(s)$ is known as sensitivity while the ratio $n(\bar{s},\bar{s})/n(\bar{s})$ is known as specificity. The prior probability $p(s)$ can be estimated as $n(s)/n$. When the system cannot classify all cases in the test set, then the confusion matrix refer to the subset of cases that were classified. Clearly, the cost analysis requires the specification of the costs $C_{s\bar{s}}$, $C_{\bar{s}s}$ and C_i or, at least, of the ratios $r_1 = C_{s\bar{s}}/C_{\bar{s}s}$ and $r_2 = C_i/r_1$.

5. Experimental Evaluation

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

5.1 Material and Methods

The 324 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

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

Table 1: The attributes of interest

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 a 5-fold cross validation experiment. 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 $\mathcal{D}_i$. 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[1] 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. The variable denoting the Glasgow Coma score was recoded into three categories representing low (≥ 4), middle ($5 - 12$) and high value (≥ 13).

For this study, a patient is classified as not surviving in hospital if his 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 greater than 0.5. The robust Bayes classifier under the complete-admissible score predicts the patient outcome as that one corresponding to the probability interval with the largest mid-point.

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

	Logistic		RBC_u		RBC_d			
	Predicted		Predicted		Predicted			
True	s	$\bar{s}$	s	$\bar{s}$	s	$\bar{s}$	Fail	Total
s	210	22	202	30	170	13	49	232
$\bar{s}$	42	50	39	53	26	39	27	94

Table 2: Confusion matrix for the three classification methods. s denotes survival while $\bar{s}$ denotes not survival. Fail refers to the cases that were not classified by the RBC_d.

As each data set D_i contains the observed outcome, we evaluated the performance of the two models by comparing their estimated accuracy and coverage, and by making a cost analysis. The estimated accuracy is 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 complete-admissible score. The coverage of the robust Bayes classifier that uses the strong dominance score is the ratio between the number of cases that were classified and the size of the data set. We also provide 95% confidence limits for these figures, based on an asymptotic approximation of a Binomial distribution.

5.2 Results

Table 2 displays the confusion matrix for the prediction based on logistic regression, on the RBC_d and on the RBC_u. 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 strong dominance criterion increases to 84.25%±2.05. The price of such increased accuracy is a decreased coverage of 76.54%±2.06. Using the complete-admissible score, we increased the coverage of the robust Bayes classifier to 100% by reducing the accuracy to 78.70%±2.15, which is inferior to the accuracy achieved by logistic regression. Both specificity and sensitivity of the RBC_d (92.90% and 60.00%) are higher than those of the classification based on logistic regression (90.52% and 54.35%) and the RBC_u (87.06% and 57.61%). However, the RBC_d is unable to predict the outcome of 21.12% of patients who survived while at the hospital and 29.35% of patients who died. A summary of these results is in Table 3.

Although the overall accuracy of the RBC_u is inferior to that achieved with logistic regression, it is interesting to note that the RBC_u made better predictions on the outcome of the patients who died while at the hospital. Hence, either a cost-free comparison of logistic regression with the RBC_u or a comparison based on the assumption that the costs $C_{s\bar{s}}$ and $C_{\bar{s}s}$ are equal results in choosing logistic regression as the best prognostic system. However, if we, more realistically, assume that the costs $C_{s\bar{s}}$ and $C_{\bar{s}s}$ are different, the comparison of logistic regression with the RBC_u depends on the comparison

of the expected costs that are

$$\begin{aligned} C(\text{RBC}_u) &= C_{s\bar{s}}0.1294 \times 0.716 + C_{\bar{s}s}0.4240 \times 0.284 \\ C(\text{Logistic}) &= C_{s\bar{s}}0.0948 \times 0.716 + C_{\bar{s}s}0.4565 \times 0.284 \end{aligned}$$

and the RBC_u yields a smaller cost if $C_{\bar{s}s} \geq 4347.83C_{s\bar{s}}$. Here, we have used prior probability $p(s) = 0.716$ as deduced from the confusion matrix 2. The comparison of logistic regression and the RBC_u with the RBC_d needs to take into account the cost C_i incurred in not being able to give a machine-based prediction. This overall cost is computed as

$$C(\text{RBC}_d) = \{C_{s\bar{s}}0.071 \times 0.738 + C_{\bar{s}s}0.4 \times 0.262\}0.7654 + 0.2346C_i$$

where $0.738 = p(s)$ as deduced from the subset of cases classified by the RBC_d. If we suppose that the costs $C_{s\bar{s}}$ and $C_{\bar{s}s}$ are equal to C_w, logistic regression gives better prediction than the RBC_u so that the choice of the system with minimum costs is limited between the RBC_d and logistic regression. The RBC_d yields smaller costs than logistic regression if $C_w \geq 3.05C_i$ and, hence, we evaluate the cost of a wrong classification to be at least 3.05 times bigger the cost of not being able to give an automatic prediction. The comparison between the three systems becomes more complex when the costs $C_{s\bar{s}}$ and $C_{\bar{s}s}$ are supposed to be different. The cost analysis leads to choose the RBC_d whenever

$$C(\text{RBC}_d) \leq \min\{C(\text{RBC}_u); C(\text{Logistic})\}$$

and hence whenever

$$0.041C_{s\bar{s}} + 0.08C_{\bar{s}s} + 0.2346C_i \leq \min\{0.0927C_{s\bar{s}} + 0.1204C_{\bar{s}s}; 0.0679C_{s\bar{s}} + 0.1296C_{\bar{s}s}\}.$$

For example, if $C_{\bar{s}s} = 5,000C_{s\bar{s}}$, so that the RBC_u yields smaller costs than logistic regression, then the comparison between the RBC_u and the RBC_d leads to choose the RBC_d whenever $C(\text{RBC}_d) \leq C(\text{RBC}_u)$ which holds when $400.01C_{s\bar{s}} + 0.2346C_i \leq 602.0927C_{s\bar{s}}$, and therefore when $C_{s\bar{s}} \geq 0.0012C_i$. On the other hand, an evaluation $C_{\bar{s}s} = 1,000C_{s\bar{s}}$ implies that logistic regression achieves higher accuracy than the RBC_u with smaller costs. Therefore, the choice reduces to compare the RBC_d with logistic regression and the former is to be preferred whenever $80.041C_{s\bar{s}} + 0.2346C_i \leq 129.668C_{s\bar{s}}$ and, hence, when $C_{s\bar{s}} \geq 0.005C_i$.

	Accuracy θ	Sensitivity θ_{ss}	Specificity $\theta_{\bar{s}\bar{s}}$	Coverage γ
RBC_d	0.8425	0.9290	0.6000	0.7654
RBC_u	0.7870	0.8706	0.5761	1.0000
Logistic	0.8025	0.9052	0.5435	1.0000

Table 3: Estimates of the overall accuracy, sensitivity and specificity, and coverage in the three models.

One point worth noting is that, in [13], the difference in accuracy achieved by the RBC$_d$ and the RBC$_u$ compared to the standard NBC that works under the missing at random assumption was used to evaluate the impact that such assumption has on the classification accuracy. Here, since we are comparing different models with different ways of treating the missing data, the different accuracy values cannot be attributed to the assumption made by logistic regression that data are missing completely at random. To evaluate the effect of this assumption we computed the accuracy, sensitivity and specificity of the standard NBC with the same data sets and the values achieved are $\hat{\theta} = 0.7685$, $\theta_{ss} = 0.8534$ and $\theta_{\bar{s}\bar{s}} = 0.5543$. Compared to the same values achieved by the RBC$_u$ and the RBC$_d$, these figures are all smaller and support the hypothesis that the missing completely at random assumption reduces the classification accuracy, probably because the real missing data mechanism is more complex.

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 increases 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. A careful cost analysis also helps the choice of the prognostic system by taking into account different consequences of wrong predictions.

The fact that, even under strong dominance, the accuracy is limited to 84.25%±2.05 questions the ability of the models considered to represent the real dependence of the patient outcome on the 10 attributes recorded in the data set. 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. Another aspect that needs further investigation is the accuracy of the measurement of the Glasgow Coma Score, as noted already by [11].

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: Results 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] D. J. Hand. *Construction and Assessment of Classification Rules*. Wiley, New York, 1997.

[5] 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.

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

[7] 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.

[8] 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.

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

[10] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

[11] A. McQuatt, P. J. D. Andrews, D. H. Sleeman, V. Corruble, and P. A. Jones. The analyses of head injury data using decision tree techniques. In *Proceedings of of the Artificial Intelligence in Medicine, IMDM'99*, pages 336–345, Heidelberg, 1999. Springer Verlag.

[12] 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.

[13] 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.

[14] M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, 2000. To appear. Also available at http://kmi.open.ac.uk/techreports/KMi-TR-28.

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

[16] D. B. Rubin. *Multiple Imputation for Nonresponse in Survey*. Wiley, New York, 1987.

[17] 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.