A Comparison of Logistic Regression to Decision Tree Induction in the Diagnosis of Carpal Tunnel Syndrome

Stephan M. Rudolfer
Mathematics Department
University of Manchester, England

Georgios Paliouras
Informatics and Telecommunications Institute
National Center for Scientific Research, Greece

Ian S. Peers
Human Communications Science Department
University of Sheffield, England

ABSTRACT

This paper aims to compare and contrast two types of model (logistic regression and decision tree induction) for the diagnosis of carpal tunnel syndrome using four ordered classification categories. Initially, we present the classification performance results based on more than two covariates (multivariate case). Our results suggest that there is no significant difference between the two methods. Further to this investigation, we present a detailed comparison of the structure of bivariate versions of the models. The first surprising result of this analysis is that the classification accuracy of the bivariate models is slightly higher than that of the multivariate ones. In addition, the bivariate models lend themselves to graphical analysis, where the corresponding decision regions can easily be represented in the two-dimensional covariate space. This analysis reveals important structural differences between the two models.

1 INTRODUCTION

In recent years, the family of methods suitable for classification problems has been extended to include a range of new techniques, such as neural networks and decision tree induction. This observation has led to an increase in the number of empirical comparisons of classification methods on a variety of problems. The European StatLog project [19] can be considered the epitome of this work, comparing 24 techniques on 23 datasets. Unfortunately, the main conclusion of this project was that the performance of the techniques, both in absolute and relative terms, varied considerably for different datasets. As a result, the choice of technique seems to be strongly dependent on the application. An interesting attempt was made in the StatLog project to characterise the classification problem in terms of some generic features of the datasets, e.g., number and types of covariates (independent variables), number of classification categories. The aim was to justify the performance of the classification techniques on the basis of the problem types. This attempt had only limited success, due to the inadequacy of the describing features and wide variability between datasets.

The work presented in this paper is much more restricted in its scope. It compares the performances of two classification techniques on a medical problem arising in electromyography (EMG), with the aims of investigating which variables are important for classification and examining any interesting features of the dataset. The comparison uses the diagnosis of an experienced electromyographer as "gold standard". The techniques which are compared are: *decision tree induction* (DT) and *logistic regression* (LR). The medical problem on which they are compared is Carpal Tunnel Syndrome (CTS), a cluster of hand symptoms caused by entrapment of the median nerve at the wrist; see, for example, [23]. CTS is the most commonly seen nerve entrapment syndrome in hospital EMG clinics. This lends itself naturally to statistical modelling of CTS. The first author collaborated with the late Dr. John L. James, Consultant Physician, St. Luke's Hospital, Huddersfield, England, on statistical approaches to CTS diagnosis using nerve conduction studies, and systematically obtained data relevant to this problem. Initially, the diagnostic classes considered were NAD (No Abnormality Detected) and CTS. A computer program (CTSS) based upon binary logistic regression was used for some years in Dr. James' EMG clinics to screen referred hands into NAD or CTS [24, 25].

However, it became apparent that the binary diagnosis was too blunt a classification, since the treatment subsequently applied (typically, splinting, steroid injection into the wrist, or operation to free the median nerve in the carpal tunnel) depends, to a large extent, on the assessed severity of the CTS. It should be pointed out, however, that Dr. James was not responsible for the treatment given to the patient. Patients were referred to him for investigation, and his usual reply letter contained a phrase of the form "These findings are consistent with ...", leaving the choice of treatment to the referring doctor. Accordingly, a multigroup classification was adopted, and a large, new and much more detailed dataset was obtained. This dataset not only contained five diagnostic classes as determined by Dr. James (NAD, mild CTS, moderate CTS, severe CTS and non-CTS abnormality - possibly also involving some degree of CTS), but also recorded the patients' histories and clinical examinations in addition to their nerve conduction studies. This new dataset provided the springboard for our comparison of the LR and DT methodologies. Initial results of this work were presented at the 6th Biennial Conference of the European Society for Medical Decision Making, 16-18 June, 1996.

One major aspect of the study is variable selection. The data contain 38 covariates in total, making a reduction in their number essential.

Another comparison of LR and DT for the binary case has been performed in a different medical domain: the diagnosis of acute cardiac ischemia. The results of this work are presented in [17], and the general conclusion was that LR performed better than DT. This argument was based on a comparison of the ROC curves for the two methods [2, 9]. In order to acquire the ROC curve for the DT, a method of obtaining rough estimates of diagnostic probabilities was employed. Probabilistic classification is unnatural for DT, and this method of constructing pseudoprobabilities has been criticised in [22].

This paper adopts an alternative, multigroup approach, considering a subset of Dr. James' dataset mentioned above. Eliminating the non-CTS abnormalities, which were relatively rare in this dataset, we are left with four ordered diagnostic categories: NAD, mild CTS, moderate CTS and severe CTS. The LR model fitted to multicategory data differs significantly from the binary classification case, in that various approaches to the logistic modelling may be adopted [26]. Since the four categories are ordered, the most appropriate model type is the Proportional Odds (PO) model, provided its assumptions are valid for the dataset considered. The PO model was fitted using the $LOGISTIC$ procedure of the SAS statistical package [28]. For the DT, the popular $C4.5$ program [22] was used.

The main measure of comparison for the two methods is classification performance on test data independent of the design dataset, as carried out in [17] and [19]. Percentage correct classification (agreement with Dr. James' diagnosis) is the overall measure of performance used; crosstabulations of the computed diagnoses with Dr. James' diagnoses give a more detailed picture of the methods' strengths and weaknesses. However, in addition to this evaluation, an interpretation of the classification models is sought. This is done by graphical illustration of the models in the covariate space. For both DT and LR, the classification regions are bounded by hyperplanes. For bivariate models, these can easily be represented, and the areas of agreement between DT and LR mapped out. Such an analysis was made in [16], where it helped significantly in understanding the behaviour of the models. In the work presented here it has led to equally interesting results.

Section 2 of the paper presents the data that have been used in the work.

Section 3 describes the LR and DT methods for multicategory classification. Section 4 presents and discusses the results of comparing the two methods on the CTS diagnosis problem, while section 5 summarises the presented work and suggests promising extensions to it. The Appendix contains full details of the covariates used, as well as summaries of their main properties.

2 DATA FOR THE COMPARISON

The data for this comparison were collected at the Electromyography clinics of the late Dr. John L. James, Consultant Physician, St. Luke's Hospital, Huddersfield, using a pro-forma prepared by the first author in conjunction with Dr. James. The patients in the database were taken from those referred to Dr. James with queried carpal tunnel syndrome (CTS). There were 937 patients seen from March 1991 to March 1994, and the final diagnoses were divided by Dr. James into five categories: NAD, mild CTS, moderate CTS, severe CTS and non-CTS abnormality (which could include some degree of CTS). The last class was a very diffuse one, so was omitted from this study. A further reason for doing so is that the four remaining diagnostic categories are now *ordered.* The remaining patients were converted into 1710 hands, which were randomly divided into a design set of 850 hands and a test set of 860 hands. Both the decision tree and logistic regression were designed on the former and tested on the latter. The diagnostic distributions in the two subsets were kept as close as possible to that in the original set.

The variables recorded for each hand fall into three groups: history, clinical examination, and nerve conduction studies (NCS). We shall give examples of each type, and leave the full list for the Appendix.

History. This covers the patient's age, handedness and sex, as well as four symptoms playing an important role in alerting the doctor to the possibility of CTS. For example, numbness and tingling in the area of the hand innervated by the median nerve. The full list of history symptom variables and their coding is given in Table 5 of the Appendix.

Clinical Examination. This covers signs and symptoms observed or elicited by the examining doctor or technician. For example, wasting and weakness of the muscles in the median nerve-innervated part of the hand. The full list and coding of clinical examination variables is given in Table 6 of the Appendix.

Nerve Conduction Studies. These involved the use of a specialised EMG machine (Medelec M6) providing electrical stimulus to, and recording the responses from, the nerves tested. The motor and sensory fibres of the median and ulnar nerves were tested at the wrist and at the elbow. The two most important nerve conduction variables in the diagnosis of CTS are the median motor latency at the wrist and the median sensory latency. The full list of nerve conduction variables is given in Table 7 of the Appendix.

Precise definition of the severity of CTS varies from clinician to clinician. It will involve a mixture of history, clinical examination and nerve conduction studies. At one extreme, [34] defined CTS severity purely in terms of the last two (mild CTS hands were symptom-free one or more days per week; moderate CTS hands had symptoms daily, awoke the patient from sleep or required modification of daytime activity to reduce the symptoms; severe CTS hands had constant numbness and/or tingling or there was thenar muscle weakness). At the other extreme, [30] defines the severity purely in terms of nerve conduction measurements: mild CTS includes prolonged median sensory latency (MSL); moderate CTS in addition involves prolongation of the median motor latency at the wrist (MMLW); severe CTS occurs when the MMLW and MSL are prolonged, with either no median sensory response or low median sensory amplitude (MSA). In between these two extremes, we have [23], who define mild CTS as CTS in which the symptoms are transient and may resolve completely, and the nerve conduction abnormalities may resolve completely or partially. They define moderate CTS as having recurrence of hand symptoms many times per week and evidence of local slowing of nerve conduction across the carpal tunnel. Severe CTS occurs according to [23] when there is clinical evidence of median nerve damage (weakness and wasting of the thenar muscles). Dr. James determined the severity of CTS in terms of clinical examination and nerve conduction studies, mainly MMLW and MSA (personal communication), although the precise way in which he combined them is probably best described as "clinical judgement". It is possible that Dr. James' lack of reliance on a patient's history may well have been due to its unreliability. For example, patients often find it hard to remember exactly how long they have suffered from pain in the index finger. An approximate algorithm DTJJ, using MMLW and MSA only and given by Dr. James to one of his technicians, is given in section 4.2.

One important aspect of the nerve conduction variables is the occurrence of a non-response in the measured nerve. This can happen if there is damage to the relevant nerve, thereby preventing conduction of nerve impulses down the nerve pathway. In general, sensory fibres are the first to be affected by nerve damage; motor fibres are thicker than sensory fibres, so are more robust to damage. Non-response does not represent a missing value in the usual sense of the term, since an attempt has been made at acquiring it. One possible approach is to regard non-response as a sub-threshold response, which cannot be detected by the EMG machine. However, modern EMG machines use such low thresholds that it can be safely assumed that there is no response for the machine to detect in the first place.

Non-responses were coded as follows: latencies and durations were coded as

99.9, being higher than any possible recorded value; this is the practical equivalent of the mathematical concept of “infinity”; amplitudes and rates were coded as 0. Such coding of non-response is intuitively appealing, since if there is no response in a nerve, then its “waveform” will be flat (have amplitude zero), take infinitely long to occur (a latency of “infinity”), and be of infinite duration. Also, the rate of transmission of electrical impulses down a nerve with non-response must be zero. If there was a non-response in either the motor latency at the wrist or at the elbow, then the corresponding rate was also coded as a non-response. It should be stressed that these values represent purely a coding of the NCS variables and do not have any inherent numerical meaning. This will affect the use and interpretation of such values. Table 8 in the Appendix gives the distribution of non-responses in the median nerve by diagnostic class, for both the design and test sets. The three median sensory variables (latency, amplitude and duration) all had non-responses together, since they are three ways of describing the same sensory waveform. Thus, it was not necessary to indicate them separately in Table 8. The general pattern is clear: non-responses occurred mainly in the severe CTS group. For the median motor measurements, non-responses occurred *only* in the severe CTS group in the design set, and in addition once in each of MMLW and MMLE in the moderate CTS group in the test set. This means that MMLW non-response is a very good predictor of severe CTS. The picture for the median sensory non-responses is not as clear cut. In the design set, 65 non-responses occurred in the mild and moderate CTS groups, while in the test set, there were 71 non-responses in *all* the other groups than the severe CTS group. As mentioned before, a possible explanation for this is that sensory nerve fibres are thinner than motor nerve fibres, hence are more vulnerable to injury. Thus, they can be damaged even in less severe CTS.

3 METHODOLOGY

3.1 Decision Trees

Decision trees have been developed by both the machine learning and statistical communities {[22, 4, 7] and [3], respectively}. In the latter context, they are known as CARTs (Classification And Regression Trees). Decision trees have often been used in medical diagnosis [17, 22, 10, 6]. Structurally, they consist of two types of node: non-terminal (intermediate) and terminal (leaf). The former correspond to questions asked about the characteristic features (covariates) of the diagnosed case. These may be factorial, e.g., “Does the patient have symptom X?”, or ordinal, e.g., “Is symptom X absent, mild, moderate or severe?”, or continuous, e.g., “What is the patient’s median motor latency at the wrist?”. For ordinal or continuous covariates, we shall only consider binary splits of the form “Is the covariate at most k?”, where k is a cutpoint. The selection of intermediate nodes involves varying k over all possible values and selecting the "best" value in the sense of the evaluation function (see below). Terminal nodes, on the other hand generate a decision/diagnosis. Diagnosis is achieved by a stepwise decision-making process, where a single question is asked each time and, depending on the answer, a different branch of the tree containing another set of questions is followed. Figure 1 presents a simple decision tree, which is, in fact, the decision tree DT1N selected by $C4.5$ on the data used in this study (see section 4.1). The root of the tree contains the first diagnostic question asked by the classifier: "Is the median motor latency at the wrist greater than 6.5 msec?". Depending on the answer for each particular case, a different branch of the tree is traversed, arriving at a decision node, which is denoted by a rectangle in Fig. 1.

One way of building decision trees is by analysing recorded diagnosed cases. These constitute, in statistical parlance, the design set. A substantial amount of work on this task has been carried out in the areas of machine learning [4, 21], statistical pattern recognition [7] and statistics [3], resulting in an abundance of methods. One of the methods, proposed by the machine learning community, is called *Top-Down Induction of Decision Trees* ($TDIDT$) and is based on the recursive partitioning of the available design, i.e., diagnosed, cases. In brief, at each recursive step, one question is selected, which discriminates "best" among different diagnoses. This question partitions the design set, and the process is repeated for each of the subsets, until they all consist of cases with a common diagnosis. The criterion used for choosing the "best" discriminating question at each recursive step, the *evaluation function*, varies between different implementations of the $TDIDT$ method.

The program used in this paper is called $C4.5$ [22] and it is an improved version of one of the most popular $TDIDT$ algorithms: $ID3$ [21]. $ID3$ uses an information-theoretic evaluation function, which is based on the minimisation of the entropy, i.e., the information content, of the unpartitioned dataset. The entropy of a set – be it the unpartitioned set or any of its subsets – is given by the formula

$$-\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}, \tag{1}$$

where k the number of categories in the problem, n_i the number of cases in the ith category and n the total number of cases in the dataset. The implementation of $ID3$ in $C4.5$ incorporates a host of useful features, including soft thresholds for numeric attributes, pruning of decision trees and extraction of optimal rule sets from trees. These features are described in detail in [22]. Two of the parameters of the program – labelled m and c – have a significant influence on the behaviour of the system. The former determines the least number of design cases to be found on a leaf node of the decision tree and the latter is a confidence factor, taken into account when the tree is being pruned. These parameters have been tuned, where needed, using 10-fold cross-validation on the design set and optimising the

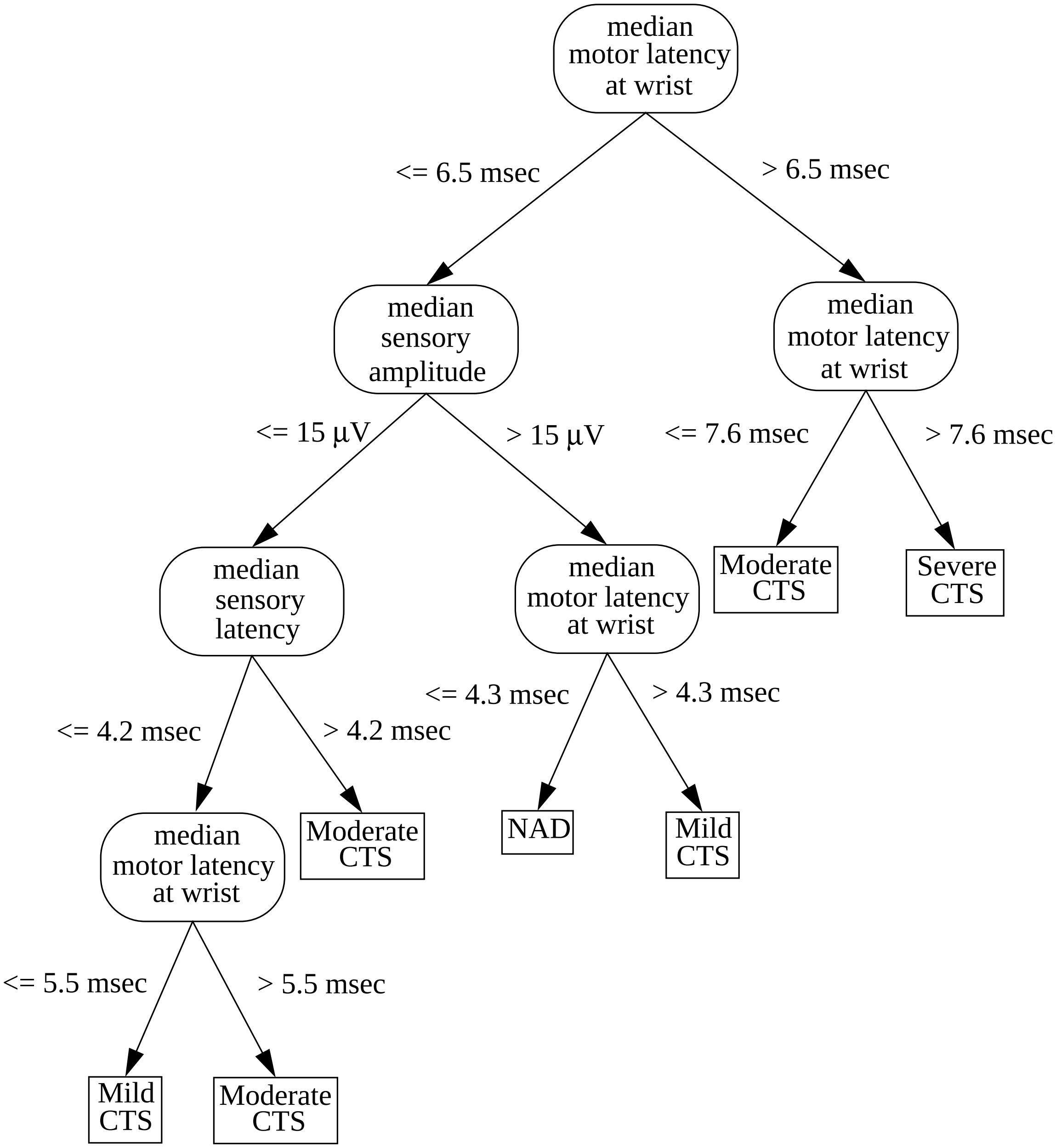

Figure 1: The multivariate decision tree DT1N.

classification accuracy, as well as the complexity, i.e., the size, of the derived decision tree.

3.2 Logistic Regression

Since the data classes are ordered, an appropriate LR model to use here is the so-called Proportional Odds (PO) model. The PO model was introduced by Walker and Duncan [33], studied in detail by McCullagh [18], and compared to other LR models by [26]. It is the most commonly used ordinal logistic regression model, and can be simply described as follows. Let Y denote the disease-class variable, where the values 1, 2, 3, 4 correspond, respectively, to severe CTS, moderate CTS, mild CTS and NAD in our context. PO assumes cutpoints $\alpha_1 < \alpha_2 < \alpha_3$ and postulates a linear form for the log odds of $Y \leq j$ $(j = 1, 2, 3)$ given the vector of covariates $\mathbf{x}$:

$$\log\{P(Y \leq j|\mathbf{x})/P(Y > j|\mathbf{x})\} \quad = \quad \alpha_j + \boldsymbol{\beta}^T\mathbf{x}, \tag{2}$$

where $\boldsymbol{\beta}$ is the vector of parameters for $\mathbf{x}$ and $P(Y \leq j|\mathbf{x})$ denotes the conditional probability of $Y \leq j$ given the covariate vector $\mathbf{x}$.

Equation {2} is termed the *proportional odds assumption*, since the ratio of the odds of the event $\{Y \leq j\}$ at $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{x} = \mathbf{x}_2$ is

$$\frac{P(Y \leq j|\mathbf{x}_1)/P(Y > j|\mathbf{x}_1)}{P(Y \leq j|\mathbf{x}_2)/P(Y > j|\mathbf{x}_2)} \quad = \quad \exp\{\boldsymbol{\beta}^T(\mathbf{x}_1 - \mathbf{x}_2)\},$$

which is independent of the choice of category j.

Equation {2} can be expressed equivalently in terms of the logits of the cumulative probabilities, where $\text{logit}(x) = \ln\{x/(1-x)\}$ for $0 < x < 1$:

$$\begin{aligned}
\text{logit}\{P(\text{severe CTS}|\mathbf{x})\} &= \alpha_1 + \boldsymbol{\beta}^T\mathbf{x} && (3)\\
\text{logit}\{P(\text{severe or moderate CTS}|\mathbf{x})\} &= \alpha_2 + \boldsymbol{\beta}^T\mathbf{x} && (4)\\
\text{logit}\{P(\text{severe, moderate or mild CTS}|\mathbf{x})\} &= \alpha_3 + \boldsymbol{\beta}^T\mathbf{x} && (5)\\
\text{logit}\{P(\text{NAD}|\mathbf{x})\} &= -\text{logit}\{P(\text{severe, moderate or mild CTS}|\mathbf{x})\} && (6)
\end{aligned}$$

The PO assumption is a very strong, if simplifying, one, which needs to be checked. Peterson & Harrell [20] proposed a score test of the PO assumption, against the general alternative hypothesis in which $\boldsymbol{\beta}$ in {2} is replaced by $\boldsymbol{\beta}_j$. However, the score test suffers from several drawbacks (Scott, Goldberg & Mayo [29]): spuriously low p-values may be produced if the samples are large or if a categorical covariate has zero cells at inner values of Y. Peterson & Harrell [20] carried out a simulation study using categorical covariates only, as a result of which they concluded that their score test often gives blatantly erroneous results when the corresponding contingency table has empty inner cells. They also state that the score test may be slightly anticonservative in the sense that the simulated sizes always exceeded the nominal significance levels. The latter statement was interpreted by Ananth & Kleinbaum [1] as the "extreme anticonservatism" of the score test, while the *SAS* Logistic Regression Examples [27] stated that this test is "very anticonservative". For these reasons, the p-values given in the *SAS* output for the present dataset (mainly of the order of 0.0001) were treated with considerable caution. Since the PO models performed very comparably to the DT ones, and used similar covariates, it was felt that they could be used without further investigation.

Automatic variable selection is a widely used, if much criticised, technique which is routinely provided in many statistical packages. The latter, however, do issue warnings against the unthinking use of such methods. For example, [27] states that "Model selection methods are exploratory. It is useful to verify the fit of the selected model on other data." The exploratory nature of variable selection in logistic regression is in part due to the fact that it involves multiple hypothesis testing. As a result, the actual significance level of the tests involved will be much higher than their nominal value. In addition, none of the test statistics used has exactly got a χ^2 distribution. This means that the usual results for Normal population distributions do not apply precisely in this context. Indeed, the exact distributions of the test statistics involved are not known, and only asymptotic χ^2 distributions are quoted [27].

Among outspoken critics of automatic variable selection, we can list [31], [12], [8] and [13]. Reference [13], p. 935, states that "It is tempting to use P-values and stepwise methods to develop a parsimonious prediction model. Besides invalidating confidence limits and causing measures of predictive accuracy such as adjusted R^2 to be optimistic, there are many other reasons not to rely on stepwise techniques (see Harrell *et al.* [12] for citations)." According to [12], p. 363, "Researchers apparently do not realize that when many predictor variables are analysed, variable screening based on statistical significance and stepwise variable selection involve multiple comparisons problems that lead to unreliable models. These methods are therefore not viable for data reduction (see Reference 17 for a condemnation of stepwise variable selection)." Their Reference 17 [8] deals, strictly speaking, only with least squares multiple linear regression (Normal) models, but Reference 17's comments (p.268) should, nevertheless, be taken seriously: " In short, stepwise methods were not designed to find 'best' models or to indicate the relative importance of variables. On the contrary, they were designed to select subsets from data sets 'padded with extraneous variables - for example, those that contain everything we could measure' (Hoerl *et al.* [14], p. 378)." Finally, Steyerberg *et al.* [31] point out four drawbacks to stepwise selection.

1. The selection is unstable, in that addition or deletion of a few patients in

the data set can substantially alter the selection [11].

2. Stepwise selection may have limited power to select important covariates in small data sets.
3. There is a substantial risk that one or more (almost) random covariates are selected, since multiple comparisons are made [8].
4. The coefficients of the selected covariates are biased to more extreme values. Steyerberg *et al.* [31] distinguish three types of bias.
    (a) **Selection bias** caused by selective inclusion of covariates with more extreme coefficients.
    (b) **Confounding bias** caused by correlation between selected and unselected covariates.
    (c) **Estimation bias**, a statistical artifact of regression analysis which leads to overestimation of the coefficients. Estimation bias occurs especially when many covariates are fitted in relatively small data sets [5], [32].

Bearing all the above points in mind, we have nevertheless used variable selection since firstly our dataset is large, thus nullifying objections 2. and 4.(c), secondly the covariates selected by LR are very similar to those selected by the more rigorous, 10-fold cross-validated DT, and thirdly the covariates selected by LR are clinically important.

The *SAS* LOGISTIC procedure [28] uses iteratively reweighted least squares to compute estimates of the parameters in the PO model, and outputs the result (test statistic and p-value) of the Peterson and Harrell score test. Variable selection can be done using forward selection, backward elimination or stepwise selection. For our dataset, the results obtained from these three methods were almost identical. For this reason, we have only quoted them for forward selection. As part of its output, the *SAS* LOGISTIC procedure prints out the results of the Wald tests of significance of the covariates' coefficients. Let $\hat{\beta}_j$ denote the maximum likelihood estimate of the coefficient of the jth covariate, and $\widehat{s.e.}(\hat{\beta}_j)$, its estimated standard error. The Wald χ^2 statistic is the square of the ratio $\hat{\beta}_j/\widehat{\mathrm{s.e.}}(\hat{\beta}_j)$. Under general conditions, and with large samples, this ratio is approximately distributed as a standard Normal variable. Thus, the Wald χ^2 statistic is approximately distributed as a χ^2 variable with one degree of freedom, and can be used to give a preliminary indication of the importance of the covariate. The usual remarks regarding multiple testing apply in this situation, too, which is why the Wald test is only regarded as a preliminary indicator [15].

Logistic regression itself is only a method of estimating the probabilities of the various diagnostic classes. In order to achieve a classification (diagnosis), it is necessary to specify a diagnostic algorithm. The simplest, most commonly used one, and the one used in this paper, is the highest class probability algorithm, i.e., assign to the class with the highest probability. In the unlikely case of a tied maximum probability, assignment is to the higher class, i.e., less serious diagnosis in our context.

4 RESULTS

This section presents the performance results for the two methods in the context of CTS diagnosis. Section 4.1 examines the classification performance, against Dr. James' full diagnosis, of multivariate models (with three or more covariates), which have been constructed by automatic variable selection. DT performs variable selection using the entropy metric of Eq. {1} and pruning of the generated decision tree, as explained in [22]. For the LR model, forward variable selection has been used [28], bearing in mind the reservations expressed in section 3.2. Section 4.2 reduces the dimensionality of the problem by choosing two variables which are of particular diagnostic importance. It also includes the bivariate model DTJJ suggested by Dr. James and which approximates the nerve conduction studies part of his diagnostic approach, ignoring history and clinical signs. DTJJ, as well as the bivariate models selected by DT and LR, are analysed graphically by examining the way in which they partition the covariate space.

4.1 Multivariate Models

The pro-formas used to obtain the data were intended as an epidemiological exercise, including variables which were suspected to have some involvement in the development of CTS. As a result, a large number of variables were recorded. One of the aims of our work was to investigate which of these variables were important in the diagnosis of CTS. For this reason, two types of experiment were done. The first used all the recorded variables, i.e., history, clinical examination and nerve conduction studies. The second focused on the variables which are considered most important by EMG experts for the final diagnosis of CTS: nerve conduction studies.

However, the variable selection process on the large dataset of the first experiment eliminated almost all of the history and clinical examination variables, confirming their relatively low diagnostic significance, *in the context of the models considered.* Furthermore, comparing the performance results of both the LR and DT models on the two experiments, it became obvious that even the few history and clinical examination variables which survived the variable elimination process did not add to the diagnostic power of the selected models. As a result, these variables have been ignored in the rest of the work. This decision should be interpreted with caution, because it does not imply that the history and clinical examination variables are irrelevant for the problem. Indeed, as was discussed in section 2, this information is used by the medical experts to various degrees, and in some cases it will be given higher weight than the nerve conduction studies. However, the final diagnosis of (the cause of) carpal tunnel syndrome, namely entrapment of the median nerve at the wrist, can only be achieved with nerve conduction studies.

Thus, using the reduced dataset which contains the nerve conduction studies only, the selected decision tree is the one shown in Fig. 1, named DT1N. The chosen variables for the decision tree are: median motor latency at the wrist (MMLW), median sensory amplitude (MSA) and median sensory latency (MSL). The first interesting result is that the model is very simple and uses just three variables. Due to the recursive application of the entropy evaluation function for the selection of diagnostic variables, the variables appearing in the nodes closer to the root of the tree are more important than the ones closer to the leaves. Thus, it can be said that according to $C4.5$, MMLW is the most important discriminating variable, MSA comes second and MSL is the least important of the selected variables. This result agrees with the use of the variables by some medical experts, who base their diagnosis of CTS mainly on the MMLW and MSA measurements. Others, however, regard MSL as the most important criterion for diagnosing CTS, since MSAs reflect only secondary axonal damage and their normal limits are more variable than those of MSL (Dr. K. M. Spillane, personal communication).

The model selected by LR (LRM0) uses a similar set of variables to that used by DT1N: MMLW, median motor latency at the elbow (MMLE), MSA and ulnar motor latency at the elbow (UMLE). The important variables MMLW and MSA are again selected, plus two latency measurements at the elbow. Linearity and additivity of the LR model were checked by fitting second-order, logarithmic and interaction terms involving MMLW, MSA, MMLE and UMLE. However, inclusion of the extra terms did not improve the model's performance, and they were therefore omitted. The relative importance of the variables MMLW and MSA is reflected in the values of their Wald χ^2 statistics, which exceed the corresponding values of the other covariates by factors of at least ten. The positive sign of the coefficients of MMLW and MMLE (see table 1) indicate that increased MMLW or MMLE raise the probability of CTS, at all levels of severity, while the negative signs of the coefficients of MSA and UMLE indicate the reverse. This follows from Eq. $\{3\}$ – $\{6\}$, and is consistent with the known neurophysiological fact that damage to the median nerve at the wrist in general results in longer median motor latency at the wrist and elbow, as well as lower median sensory amplitude. The inclusion of UMLE among the chosen covariates illustrates the points made in section 3.2. However, since LRM0's performance is very close to that of DT1N, we have continued to use it for comparison purposes. The selection of MMLW by both DT and LR models is surely related to the fact that MMLW non-response is a very good predictor of severe CTS (see section 2).

Variable	Parameter Estimate	Standard Error	Wald χ^2	Pr > χ^2
α_1	–8.1040	0.9461	73.3687	0.0001
α_2	–3.2187	0.7860	16.7692	0.0001
α_3	1.1286	0.7734	2.1295	0.1445
MMLW	1.1691	0.1041	126.1253	0.0001
MMLE	0.0792	0.0359	4.8843	0.0271
MSA	–0.2207	0.0158	194.6008	0.0001
UMLE	–0.2856	0.0802	12.6704	0.0004

Table 1: Coefficients of the LRM0 logistic regression model fitted to the design dataset.

In order to evaluate the models' performances, crosstabulations of the diagnoses of DT1N and LRM0 with Dr. James' diagnoses on the test set of unseen cases were obtained. These are shown in Table 2. The percentages correct for the two methods were 79.3% for DT1N and 78.4% for LRM0. The difference between the two percentages is well within sampling error, suggesting that the two models have similar diagnostic power. The more detailed performance results of Table 2 reinforce this suggestion. The patterns of diagnosis for the two algorithms are roughly the same. The majority of misdiagnoses are one category away from Dr. James' diagnosis, e.g., DT1N classifies 56 NAD cases as mild CTS, but only one as moderate CTS and none as severe CTS. This is an encouraging result, showing that the two models have captured the ordering of the four categories in the covariate space. In that sense, the misdiagnoses can be justified by the uncertainty in the definition of the boundaries between the four categories. For instance, the distinction between mild and moderate CTS will not always be clear, and Dr. James will have taken the clinical examination into account when deciding on the classification of cases close to the boundary. This point is illustrated further with the simpler bivariate models in section 4.2.

As a comparison, we fitted the full model involving all history, clinical examination and nerve conduction studies variables. Its performance was disappointing, achieving only 75.1% correct classifications.

4.2 Bivariate Models

Dr. James had designed a simple decision rule, using MMLW and MSA, which we shall denote DTJJ. This rule – as described to the first author by one of Dr. James' technicians – is as follows:

```
    IF MSA $= 0$ OR MMLW $= 99.9$ THEN Severe CTS
    ELSE IF MSA $\leq 15$ $\mu$V AND MMLW $> 4.6$ msec THEN Moderate CTS
```

(a) Decision Tree DT1N

Dr. James' Diagnosis	DT1N Diagnosis				
Frequency	NAD	Mild CTS	Moderate CTS	Severe CTS	Total
NAD	318	56	1	0	375
Mild CTS	29	247	20	1	297
Moderate CTS	1	42	90	8	141
Severe CTS	0	1	19	27	47
Total	366	247	134	113	860

(b) Logistic Regression LRM0

Dr. James' Diagnosis	LRM0 Diagnosis				
Frequency	NAD	Mild CTS	Moderate CTS	Severe CTS	Total
NAD	314	61	0	0	375
Mild CTS	33	239	25	0	297
Moderate CTS	2	36	97	6	141
Severe CTS	1	2	20	24	47
Total	350	338	142	30	860

Table 2: Crosstabulation of the multivariate decision tree and logistic regression algorithms by Dr. James' diagnosis.

```
ELSE IF MSA $\leq 15\ \mu$V OR MMLW $> 4.6$ msec THEN Mild CTS
ELSE NAD
```

The first case in the above simple rule deals with severe CTS diagnoses, which are regarded by the rule as the only ones without median sensory response or median motor response at the wrist. This reflects the fact, noted in section 2, that MMLW, and to a lesser extent MSA, is a very good predictor of severe CTS. Excluding these cases, the rest are separated by dichotomising each of the two variables MMLW and MSA. An interesting observation at this point is the proximity of the selected dichotomising values in DTJJ and DT1N, e.g., the MSA node in DT1N examines the MSA value 15 μV, as does DTJJ. It should be noted at this point that DTJJ performed relatively poorly against Dr. James' full diagnosis; it only achieved 70.1% agreement with the latter. This is consistent with the fact that Dr. James also took the clinical examination into account when reaching his diagnosis.

The simplicity of the DT1N and LRM0 models and Dr. James' bivariate decision rule DTJJ raised an interesting question: how would bivariate DT and LR models perform? Thus, a further experiment was done, generating the models BVDT (bivariate DT) and BVLR, which use only two independent variables. The natural choice of variables is MMLW and MSA, since they are used in DTJJ, DT1N and LRM0. Checks for linearity and additivity in BVLR were carried out; no reason was found to reject them. Figure 2 presents the simple BVDT and Table 3 contains the coefficients for BVLR.

	Parameter	Standard		
Variable	Estimate	Error	Wald χ^2	Pr > χ^2
α_1	–9.3290	0.7492	155.0470	0.0001
α_2	–4.7364	0.5169	83.9695	0.0001
α_3	–0.4739	0.4744	0.9980	0.3178
MMLW	1.1794	0.0970	147.9140	0.0001
MSA	–0.2182	0.0155	197.7638	0.0001

Table 3: Coefficients of BVLR, the bivariate logistic regression model fitted to the design dataset.

The simplicity of the bivariate models is very appealing. However, one needs to question their diagnostic accuracy. Intuition suggests that the classification performance of the bivariate models will be lower than that of the multivariate ones. After all, the variable selection methods could have reduced the DT1N and LRM0 models to be bivariate ones, but they did not. Surprisingly, the intuitive outcome is not verified in this case. The performance of the two bivariate models is slightly *higher* than that of their multivariate counterparts. Table 4 presents the results for the BVDT and BVLR models. The percentages correct were 80.6%

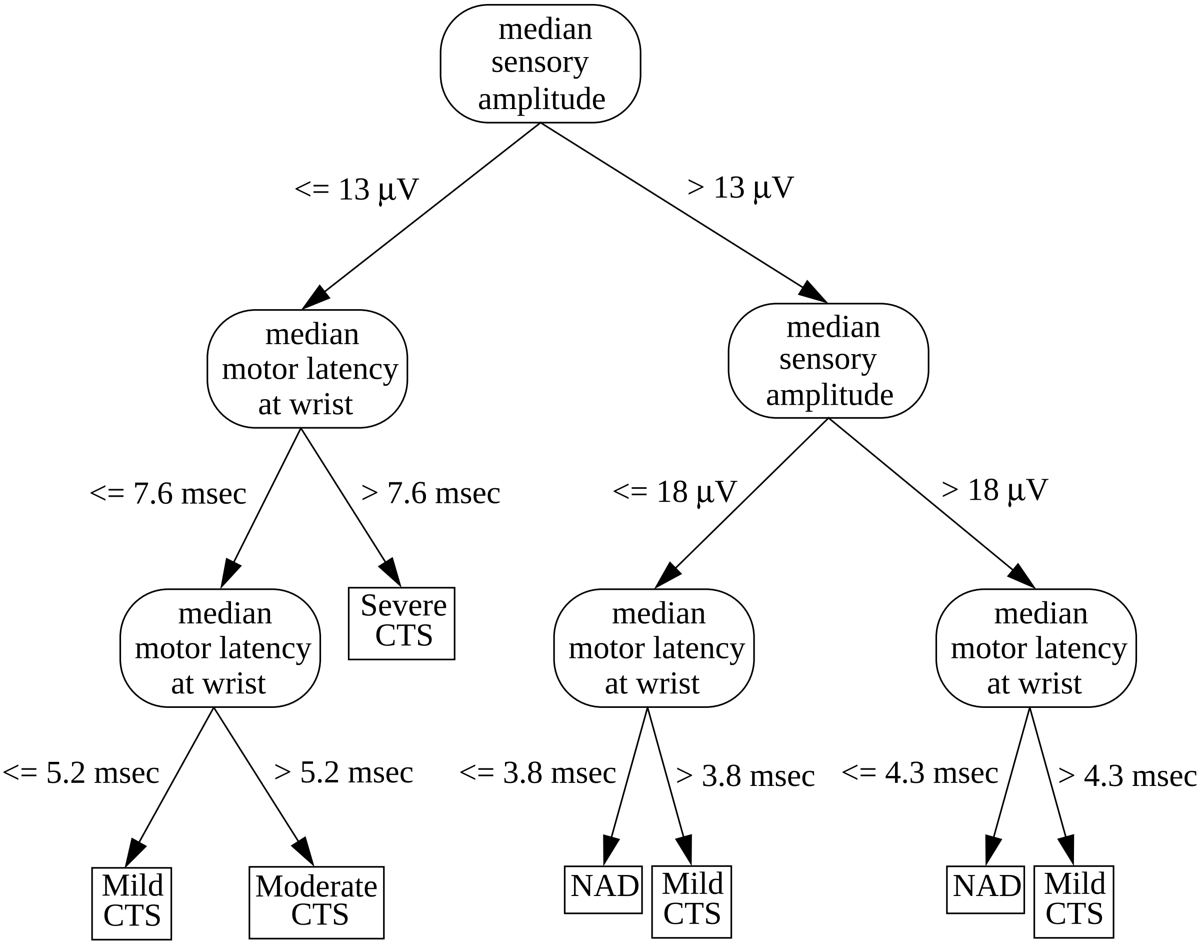

Figure 2: The bivariate decision tree BVDT.

for BVDT and 79.4% for BVLR, an increase over the multivariate models of 1% for LR and 1.3% for DT, well within sampling error. The more detailed picture drawn by the crosstabulation results is broadly similar to that for the multivariate case.

(a) Decision Tree BVDT

Dr. James' Diagnosis	BVDT Diagnosis				
Frequency	NAD	Mild CTS	Moderate CTS	Severe CTS	Total
NAD	348	27	1	0	375
Mild CTS	36	239	22	0	297
Moderate CTS	1	54	79	7	141
Severe CTS	1	8	11	27	47
Total	386	328	112	34	860

(b) Logistic Regression BVLR

Dr. James' Diagnosis	BVLR Diagnosis				
Frequency	NAD	Mild CTS	Moderate CTS	Severe CTS	Total
NAD	318	57	0	0	375
Mild CTS	33	239	25	0	297
Moderate CTS	2	34	99	6	141
Severe CTS	1	2	17	27	47
Total	354	332	141	33	860

Table 4: Crosstabulation of the bivariate decision tree and logistic regression algorithms by Dr. James' diagnosis.

One possible explanation for the higher classification accuracy of the bivariate models could be that the multivariate models overfit the design data. This is very unlikely, however, due to the simplicity of DT1N and LRM0. Moreover, the classification performance on the design set is almost identical for the multivariate and the bivariate DT models. DT1N's performance is 79.4% correct on the design set, while BVDT's is 79.3%. On the test set, DT1N's performance remains almost unchanged at 79.3%, while there is a small improvement for BVDT to 80.6% correct. The multivariate LR, LRM0, performs roughly the same on the design (78.5% correct) as on the test set (78.4% correct), while the bivariate model, BVLR, performs substantially worse on the design set (76.9% correct) than it does on the test set (79.4% correct). This latter result is perhaps surprising, since it is well known that resubstitution error rates tend to be lower than error rates on test sets. Perhaps there are some hidden features of the design set which prevent the bivariate LR model from performing better than its multivariate counterpart. Thus, the bivariate models are truly better in the test set than the multivariate ones. The failure of the variable selection methods to discover this characteristic of the problem should be interpreted as a deficiency of their search biases. Both the forward selection method for LR and the recursive partitioning for DT perform a greedy search in the space of possible variable sets and therefore do not guarantee the optimal solution. The unexpected result obtained here suggests that the results of the variable selection methods should be treated with caution, as indeed pointed out in [27] for logistic regression and discussed in section 3.2.

A further interesting feature of the bivariate models is that they can be represented graphically, in terms of the corresponding decision regions in the two-dimensional covariate space. Fig. 3 presents the comparison between BVDT and DTJJ in the (MSA,MMLW) space. Both models divide the space up into (possibly infinite) rectangular regions, using axis-parallel discrimination lines. It is clear from Fig. 3 that BVDT is a refined version of the simpler DTJJ rule. The hatched areas in the graph correspond to the disagreement between the two models. Their main difference is that the discriminating MSA value is 13 μV for the BVDT, instead of 15 μV for DTJJ. The additional discrimination attempted by BVDT with the use of MSA=18 μV seems to be unnecessary. Its omission should not affect the performance of the model significantly. An additional difference between BVDT and DTJJ is in the diagnosis of severe CTS. BVDT allocates a larger area to this category, corresponding to high MMLW values and low MSA. This result suggests that the simple definition of severe CTS in DTJJ, as non-response in either of the two nerves, is not a sufficiently broad one. There is a group of severe CTS cases in the design set which have responses for both nerves. Furthermore, many of the sensory non-response cases are not diagnosed by Dr. James as severe CTS (see table 8).

Figure 4 presents the decision regions for BVLR and DTJJ. Note that the decision regions for the LR are bounded by parallel sloping lines. This fits in with the definition of the PO model for LR. The lines have a positive slope, representing the fact that the two covariates are related with CTS in opposite ways, i.e., as MSA increases and MMLW decreases, the probability of CTS decreases. Moreover, the ordering of the four categories is preserved in the ordering of the decision regions, i.e., adjacent regions correspond to consecutive categories. Again the areas of disagreement between the models are indicated by various forms of hatching. The fit between the two models is not as close as between BVDT and DTJJ, because of the different nature of the two models. However, there are still large areas of agreement. Moreover, the good performance of the BVLR model,

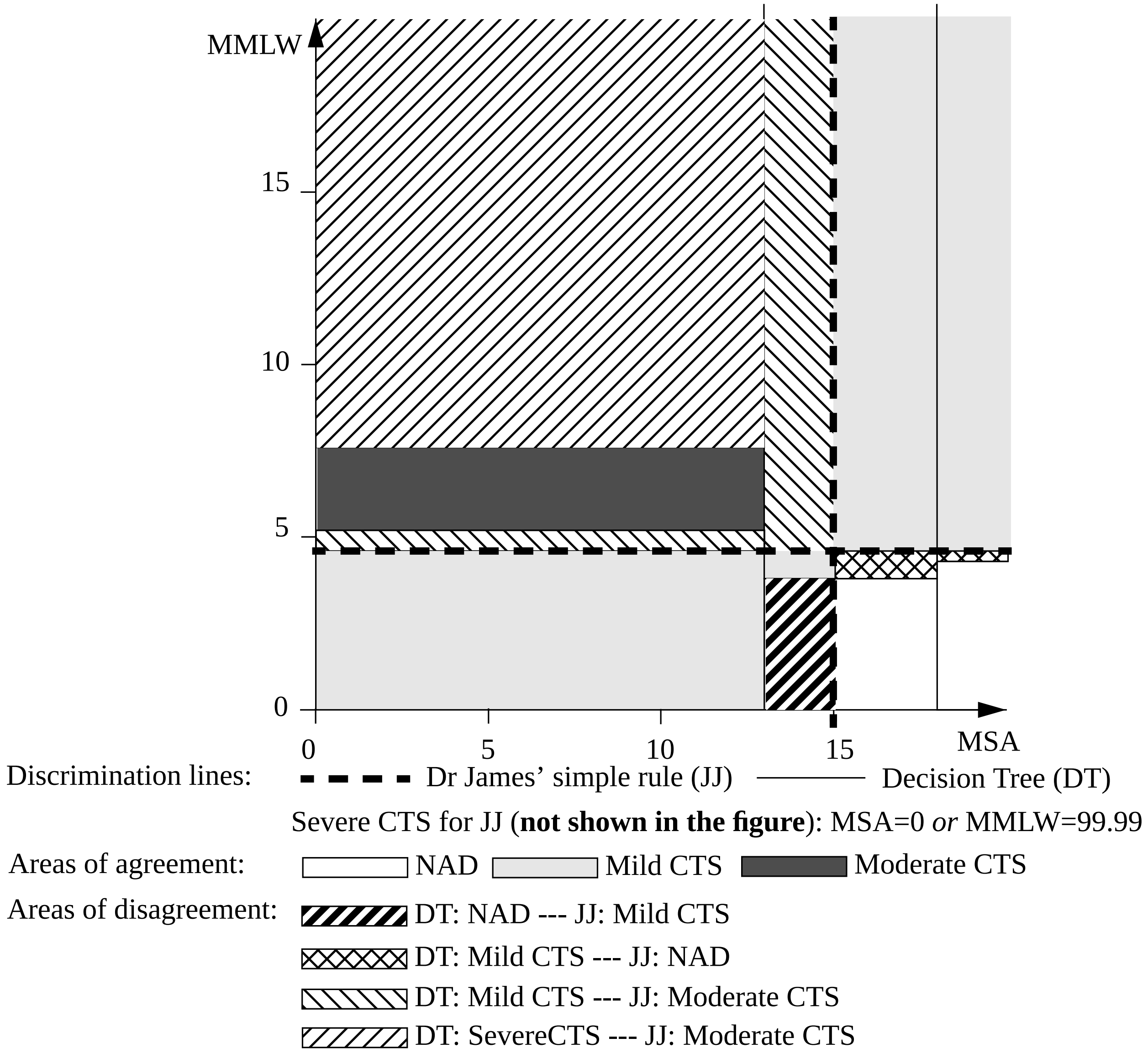

Figure 3: Comparison of Dr. James' simple rule DTJJ with the bivariate decision tree BVDT. Notation: MMLW = median motor latency at the wrist; MSA = median sensory amplitude.

suggests that the ordering imposed by the PO assumption on the decision regions is suitable for this problem. Finally, BVLR treats the severe CTS category in a similar manner as BVDT, i.e., it allocates to it an area of high MMLW values.

Figure 5 compares the BVDT and BVLR models, showing the areas where they disagree by one or two classes. Ignoring the severe CTS class, the disagreement between BVDT and BVLR is almost identical to that between BVLR and DTJJ, because BVDT is very similar to DTJJ. Regarding the severe CTS category, there is a large area of agreement for high MMLW and low MSA values. However, the two models disagree by two categories for large MMLW and large MSA. BVLR would classify such cases as severe CTS, while BVDT would assign them to the mild CTS group. This is shown by the area on the top right of the graph, which is the only area in which the two models disagree by two classes. A simple explanation for this phenomenon is that this area is very sparsely populated. The reason for this is that the two covariates are correlated to some extent. It would be very unusual for someone to have a high MSA when his/her MMLW is also high.

In summary, the three bivariate models largely agree in their decision regions. The decision tree BVDT is very close to Dr. James' simple rule DTJJ. This can be explained by the decision-theoretic background of decision trees, i.e., the fact that they were originally designed to help in organising the human decision making process. Another effect of this feature of decision trees is that they are comprehensible to humans, even in multivariate spaces (more than two covariates). On the other hand, BVLR also provides a natural solution to this problem, due to the ordering of the four categories, and in addition provides probabilities of the four diagnoses, something which decision trees do not do naturally (see section 1 and [22]).

5 CONCLUSIONS

The diagnosis of Carpal Tunnel Syndrome has been examined using two different modelling methods: induction of decision trees and logistic regression. A multi-group classification has been adopted, containing four ordered classes: NAD, mild CTS, moderate CTS and severe CTS. An initial data pruning was carried out in that two experiments were done: one with history, clinical examination and nerve conduction studies; the other with nerve conduction studies only. The conclusion reached was that history and clinical examination did not significantly improve classification performance *with these models.* Hence, for the purposes of this study, they were omitted from further consideration. Automatic variable selection for DT and LR was employed to reduce the nine nerve conduction study covariates to a more manageable number: three for DT and four for LR. Variable selection for LR is open to criticism, as discussed at length in section 3.2, but the selected model performed well, and hence was retained. The performances of

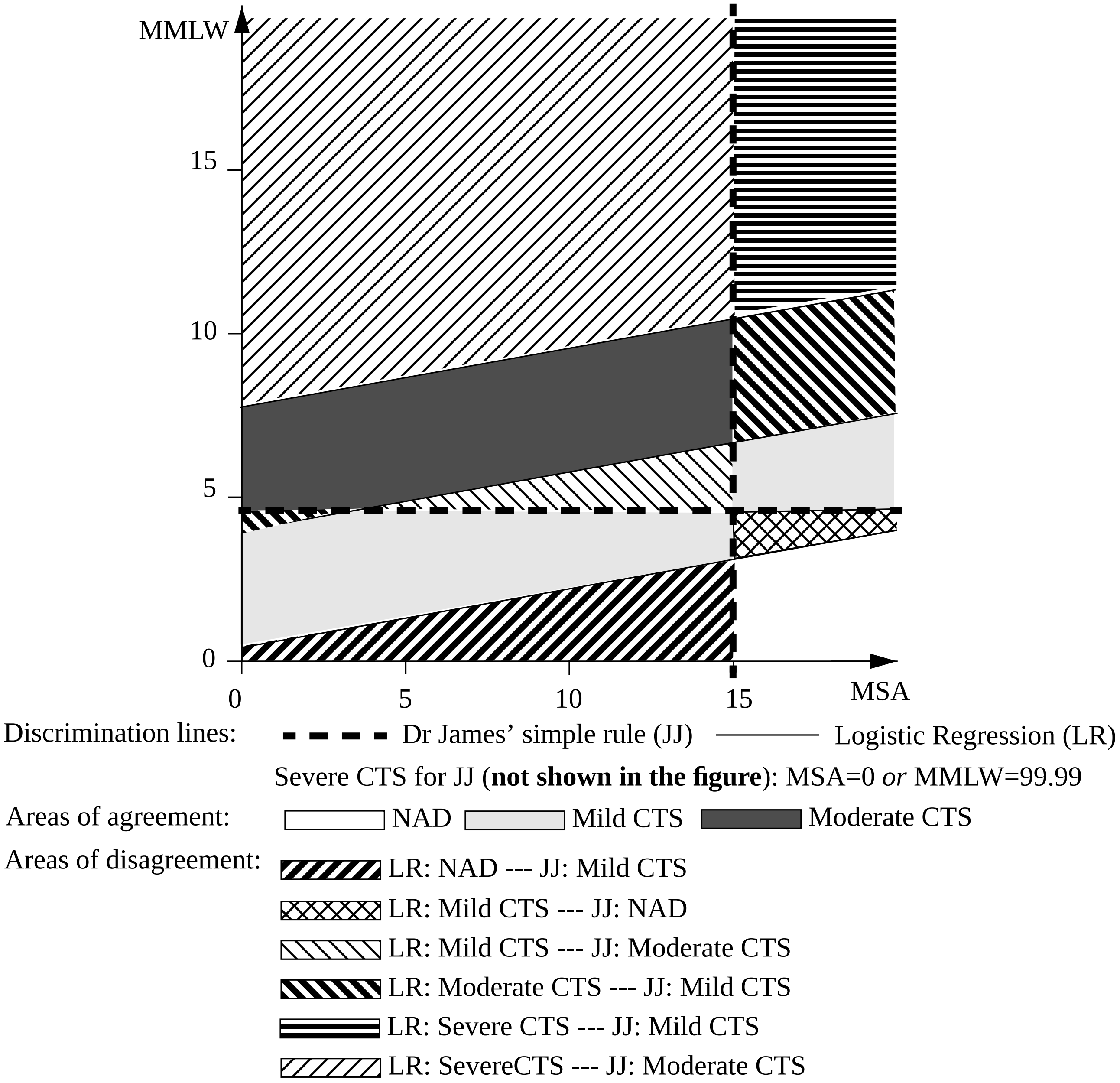

Figure 4: Comparison of Dr. James' simple rule DTJJ with the bivariate logistic regression BVLR. Notation: MMLW = median motor latency at the wrist; MSA = median sensory amplitude.

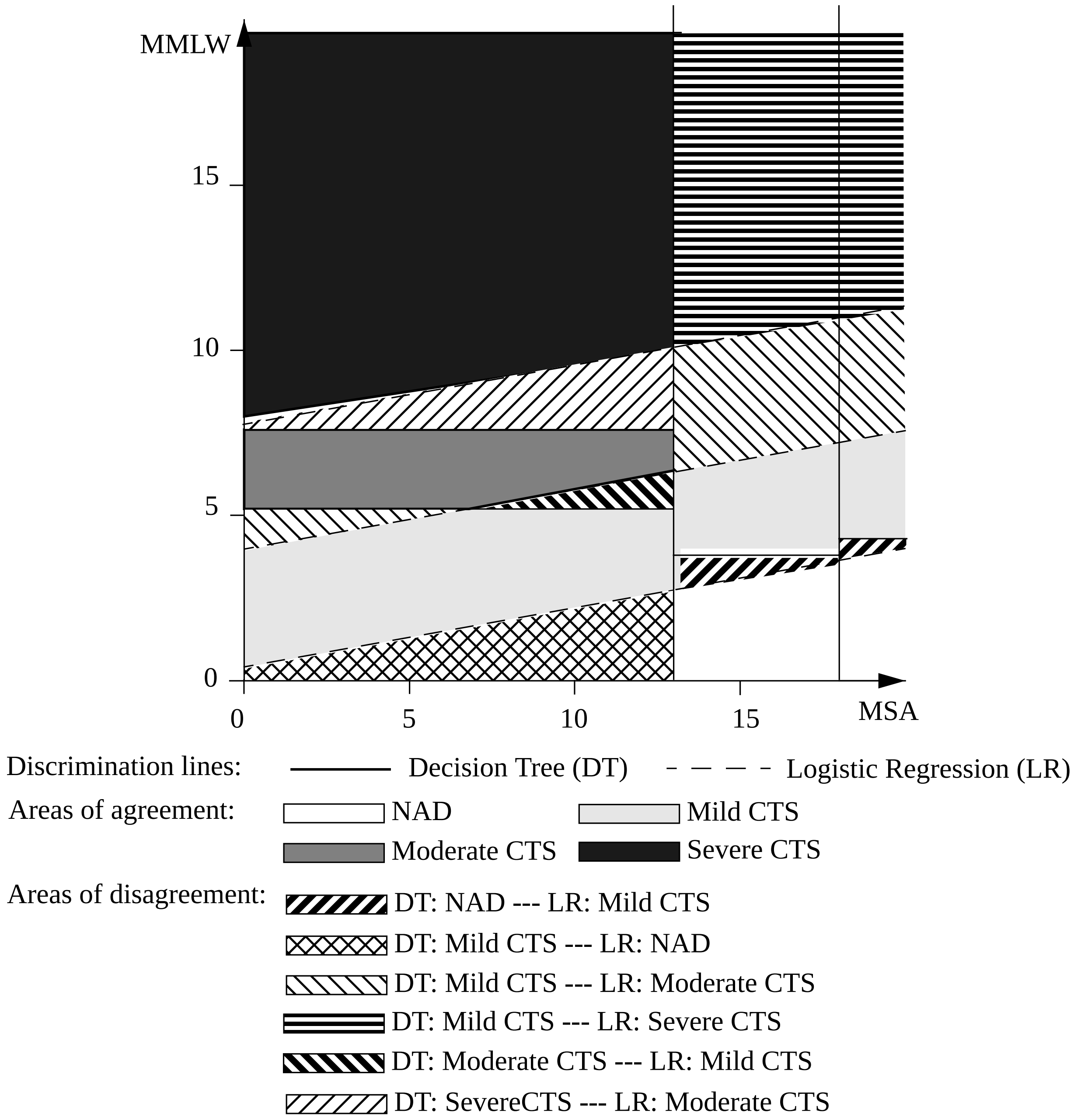

Figure 5: Comparison of the bivariate decision tree BVDT and logistic regression BVLR: regions where they differ by up to 2 classes. Notation: MMLW = median motor latency at the wrist; MSA = median sensory amplitude.

the two methods were compared by measuring the classification accuracy of the selected models, relative to Dr. James' full diagnosis, on unseen data. The two models were found to perform very similarly on this problem, achieving satisfactory classification accuracy. The coding of non-response, described in section 2, has turned out to be very effective, and both DT and LR handle this coding very well.

Furthermore, an effort was made to understand the decision process corresponding to the two models. This was achieved by examining graphically the decision regions corresponding to two bivariate models, based on the two most significant covariates. The main conclusion of this study is that, although the two models take very different approaches to the division of the covariate space, they have large areas of agreement. Especially, the bivariate decision tree has been found to agree to a large extent with a simplified decision rule due to Dr. James, which is based only on nerve conduction studies. The results of the bivariate analysis are particularly significant, due to the high classification accuracy of the bivariate models. Surprisingly, these models performed at least as well as their multivariate counterparts.

The results of the study presented here suggest that simple bivariate models are the most appropriate for the CTS dataset that has been examined, providing that attention has been restricted to nerve conduction studies variables. This conclusion holds at least for the two modelling methods participating in the study. It would be interesting to compare these results with those of a different modelling method, such as a neural network, which is capable of producing complex decision regions. In particular, it would be of interest to examine the effect of non-linear decision boundaries on the classification accuracy and the choice of covariates, possibly making use of the history and clinical examination variables. Furthermore, in view of the drawbacks to automatic variable selection, it would be interesting to compare our results with other methods of variable selection, as well as with models also incorporating history and clinical examination in a more subtle way.

Another interesting direction for further work is the combination of domain knowledge with the data analysis methods examined here. For instance, one could build a model which incorporates the results of this study, with special rules for handling exceptional cases. The aim in that case would be to provide a more accurate model of the decision process performed by the medical expert, additionally making use of the history and clinical examination variables.

A further aspect of the diagnostic problem is that of incorporating cost or loss functions, thereby taking into account the relative seriousness of the various misdiagnoses. In particular, the distance between the actual and computed diagnoses should play a part in constructing an appropriate loss function (see, for example, J. Anderson's comments on McCullagh's paper [18]).

APPENDIX

This section contains the full details of the variables used in the analysis of the CTS dataset and the distributions of the non-responses in the median nerve by diagnostic class. Table 5 presents the coding of the history symptom variables. A classical history of CTS is one of discomfort (any or all of numbness, pain, tingling, weakness) during the night, lasting typically about ten minutes and relieved by shaking the hand. The region of the hand affected is that of the median nerve (thenar eminence, thumb, first finger and thumb side of the middle finger). It was felt that the length of time the patient had experienced these symptoms might have a bearing on their condition, so was included in the questions asked.

Descriptor	Coding	
Duration	1	at most 10 minutes
	2	over 10 minutes
First	1	less than 3 months
	2	from 3 months to one year
	3	from 1 to 5 years
	4	from 6 to 10 years
	5	over 10 years
Location	1	first to third fingers
	2	fourth and fifth fingers
	3	all five fingers
	4	other
Relief	1	shaking hand
	2	other
	3	none
Severity	1	mild
	2	moderate
	3	severe
Time	1	daytime episodes
	2	nocturnal episodes
	3	episodes day and night
	4	continuous symptom

Table 5: Coding of history symptom variables: numbness, pain, tingling, weakness.

Table 6 contains three clinical examination variables (sensory loss, wasting and weakness) whose presence can indicate damage to the relevant nerve. Since these variables are elicited by the clinician or technician, it was felt that they carry a greater degree of objectivity than the symptoms described by the patient in the history symptom variables.

Variable	Coding	
Sensory Loss		
Location	1	first to third fingers
	2	fourth and fifth fingers
	3	all fingers
	4	other
Wasting		
Location	1	Thenar Eminence
	2	Hypothenar Eminence
	3	other
Severity	1	mild
	2	moderate
	3	severe
Weakness		
Location	1	Thenar Eminence
	2	Hypothenar Eminence
	3	other
Severity	1	mild
	2	moderate
	3	severe

Table 6: Coding of clinical examination variables.

Table 7 records the nerve conduction study variables used in the study. It should be noted that the ulnar measurements were included by Dr. James in order to increase the probability of detecting non-CTS abnormalities. For this reason, it would not be expected that they would play a large role in the four-class diagnostic problem considered in this paper.

Table 8 shows the distribution of the non-responses by diagnostic class. This has been discussed in detail in section 2.

Nerve	Measurement
Median	Motor Latency at the Wrist
	Motor Latency at the Elbow
	Motor Rate, Elbow to Wrist
	Sensory Latency
	Sensory Amplitude
	Sensory Duration
Ulnar	Motor Latency at the Wrist
	Motor Latency at the Elbow
	Motor Rate, Elbow to Wrist

Table 7: Nerve conduction study variables.

(a) Design Set

		NAD	Mild CTS	Moderate CTS	Severe CTS	Total
Variable	MMLW	0	0	0	11	11
	MMLE	0	0	0	12	12
	MS[a]	0	11	54	39	104
	Total	0	11	54	62	127

(b) Test Set

		NAD	Mild CTS	Moderate CTS	Severe CTS	Total
Variable	MMLW	0	0	1	4	5
	MMLE	0	0	1	4	5
	MS	1[b]	6	64	41	112
	Total	1	6	66	49	122

[a] Median Sensory Response

[b] The median sensory responses in this hand had not been recorded, but were entered as a non-response

Table 8: Distribution of non-responses by diagnostic class.

ACKNOWLEDGEMENTS

This paper is dedicated to the memory of the late Dr. John L. James, to whom the first author owes his interest in, and work on, the application of statistics to the area of electromyography, in particular carpal tunnel syndrome. The first author expresses his profound gratitude to Dr. James' secretary, Mrs. Sylvia Hague, and his technicians, Miss Gill Lockwood, Mrs. Ginny Lockwood and Mrs. Louise Mullinger, for their great patience and cooperation in the acquisition and evaluation of the data. The authors also thank the following (in alphabetical order) for their helpful comments on a draft of this paper: Prof. M. de Krom, Dr. Diederik W. J. Dippel, Mrs. Anna Hart, Mr. John Hinde, Prof. Peter Jones, Prof. Emmanuel Lesaffre, Miss Gill Lockwood, Dr. K. M. Spillane, Dr. Ewout W. Steyerberg.

References

[1] C. V. ANANTH and D.G. KLEINBAUM. Regression models for ordinal responses: a review of methods and applications. *Int. J. Epidemiology*, 26:1322–1333, 1997.

[2] J. R. BECK and E. K. SHULTZ. The use of ROC curves in test performance evaluation. *Arch. Pathol. Lab. Med.*, 110:13–20, 1986.

[3] L. BREIMAN, J. H. FRIEDMAN, R. A. OLSHEN, and C. J. STONE. *"Classification and Regression Trees"*. Wadsworth, Pacific Grove, California, 1984.

[4] G. CESTNIK, I. KONONENKO, and I. BRATKO. Assistant-86: a knowledge-elicitation tool for sophisticated users. In I. BRATKO and N. LAVRAČ, editors, *"Progress in Machine Learning"*, pages 31–45. Sigma Press, Wilmslow, UK, 1987.

[5] J. B. COPAS. Regression, prediction and shrinkage (with discussion). *J. Royal Stat. Soc., Ser. B*, 45:311–354, 1983.

[6] N. J. CRICHTON, J. P. HINDE, and J. MARCHINI. Models for diagnosing chest pain: is CART helpful? *Stat. Med.*, 16:717–727, 1997.

[7] G. R. DATTATREYA and L. N. KANAL. Decision trees in pattern recognition. In L. N. KANAL and A. ROSENFELD, editors, *"Progress in Pattern Recognition"*, pages 189–239. North-Holland, Amsterdam, 1985.

[8] S. DERKSEN and H. J. KESELMAN. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br. J. Math. Stat. Psychol.*, 45:265–282, 1992.

[9] J. P. EGAN. *"Signal Detection Theory and ROC Curve Analysis"*. Academic Press, New York, 1975.

[10] R. F. FAUBERTAS, L. E. RODEWALD, S. G. HUMISTON, and P. G. SZILAGYI. ROC Curves for classification trees. *Med. Decis. Making*, 14:169–174, 1994.

[11] F. E. HARRELL, K. L. LEE, R. M. CALIFF, D. B. PRYOR, and R. A. ROSATI. Regression modelling strategies for improved prognostic prediction. *Stat. Med.*, 3:143–152, 1984.

[12] F. E. HARRELL, K. L. LEE, and D. B. MARK. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, 15:361–387, 1996.

[13] F. E. HARRELL, P. A. MARGOLIS, S. GOVE, K. E. MASON, E. K. MULHOLLAND, D. LEHMANN, L. MUHE, S. GATCHALIAN, H. F. EICHENWALD, and THE WHO/ARI YOUNG INFANT MULTICENTRE STUDY GROUP. Development of a clinical prediction model for an ordinal outcome: The World Health Organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis and meningitis in young infants. *Stat. Med.*, 17:909–944, 1998.

[14] R. W. HOERL, J. H. SCHUENEMEYER, and A. E. HOERL. A simulation of biased estimation and subset regression techniques. *Technometrics*, 28:369–380, 1986.

[15] D. W. HOSMER and S. LEMESHOW. *"Applied Logistic Regression"*. John Wiley & Sons, New York, 1989.

[16] H. JESSEN and G. PALIOURAS. Predicting labour force participation of women with the use of statistical and learning classification techniques. In *"European Conference in Non-Linear Econometrics (EC2)"*, 1995.

[17] W. J. LONG, J. L. GRIFFITH, H. P. SELKER, and R. B. D'AGOSTINO. A comparison of logistic regression to decision-tree induction in a medical domain. *Comput. Biomed. Res.*, 26:74–97, 1993.

[18] P. MCCULLAGH. Regression models for ordinal data (with discussion). *J. Royal Stat. Soc., Ser. B*, 42:109–142, 1980.

[19] D. MICHIE, D. J. SPIEGELHALTER, and C. C. TAYLOR. *"Machine Learning, Neural and Statistical Classification"*. Ellis Horwood, New York, 1994.

[20] B. PETERSON and F. E. HARRELL. Partial proportional odds models for ordinal response variables. *Appl. Stat.*, 39:205–217, 1990.

[21] J. R. QUINLAN. Learning efficient classification procedures and their application to chess end games. In R. S. MICHALSKI, J. G. CARBONELL, and T. M. MITCHELL, editors, *"Machine Learning: An Artificial Intelligence Approach"*, pages 463–482. Morgan Kaufmann Publishers, Inc., 1983.

[22] J. R. QUINLAN. *"C4.5: Programs for Machine Learning"*. Morgan Kaufmann, San Mateo, CA, 1993.

[23] R. B. ROSENBAUM and J. L. OCHOA. *"Carpal Tunnel Syndrome and Other Disorders of the Median Nerve"*. Butterworth-Heinemann, Stoneham, MA, 1993.

[24] S. M. RUDOLFER. CTSS: An interactive microcomputer program for the clinical screening of carpal tunnel syndrome. I. Clinical aspects. *Electromyography Clin. Neurophysiol.*, 28:259–262, 1988.

[25] S. M. RUDOLFER. CTSS: An interactive microcomputer program for the clinical screening of carpal tunnel syndrome. II. Statistical and computational aspects. *Electromyography Clin. Neurophysiol.*, 30:483–489, 1992.

[26] S. M. RUDOLFER, P. C. WATSON, and E. LESAFFRE. Are ordinal models useful for classification? A revised analysis. *J. Statist. Comput. Simul.*, 52:105–132, 1995.

[27] SAS INSTITUTE. *"Logistic Regression Examples Using the SAS System"*. SAS Institute, Inc., Cary, North Carolina, 1995.

[28] SAS INSTITUTE. Chapter 16: the logistic procedure. In *"SAS/STAT Software: Changes and Enhancements through Release 6.11"*, pages 381–490. SAS Institute, Inc., Cary, North Carolina, 1996.

[29] S. C. SCOTT, M. S. GOLDBERG, and M. E. MAYO. Statistical assessment of ordinal outcomes in comparative studies. *J. Clin. Epidemiology*, 50:45–55, 1997.

[30] J. C. STEVENS. AAEM Minimonograph #26: the electrodiagnosis of carpal tunnel syndrome. *Muscle & Nerve*, 20:1477–1486, 1997.

[31] E. W. STEYERBERG, M. J. C. EIJKEMANS, and J. D. F. HABBEMA. Stepwise selection: an underestimated source of overestimation. *Submitted to J. Clin. Epidemiology*, 1997.

[32] J. C. VAN HOUWELINGEN and S. LE CESSIE. Predictive value of statistical models. *Stat. Med.*, 9:1303–1325, 1990.

[33] S. H. WALKER and D. B. DUNCAN. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:167–179, 1967.

[34] J. C. WHITE, S. R. HANSEN, and R. K. JOHNSON. A comparison of EMG procedures in the carpal tunnel syndrome with clinical-EMG correlations. *Muscle & Nerve*, 11:1177–1182, 1988.