Diagnosis of epithelial mesothelioma using tree-based regression analysis and a minimal panel of antibodies,

Running title: Tree regression analysis for mesothelioma

SONJA KLEBE*, MARKKU NURMINEN†, JAMES LEIGH‡ AND DOUGLAS W. HENDERSON*

*Department of Anatomical Pathology, Flinders Medical Centre, Adelaide, South Australia
†Finnish Institute of Occupational Health, Helsinki, and Department of Public Health, University of Helsinki, Finland,
‡Centre for Occupational and Environmental Health, School of Public Health, University of Sydney, New South Wales, Australia
^To whom correspondence should be addressed
SUMMARY
AIMS: Immunohistochemistry with panels of antibodies is standard procedure to distinguish between malignant mesothelioma and metastatic adenocarcinoma. Most studies assess only the sensitivity and specificity for single antibodies, even when the paper concludes by recommending an antibody panel. It was the aim of this study to use a novel statistical approach to identify a minimal panel of antibodies, which would make this distinction in the majority of cases.

METHODS: Two hundred consecutive cases of pleural malignancy (173 pleural mesotheliomas of epithelial type and 27 cases of secondary adenocarcinoma) were investigated using a standard panel of 12 antibodies (CAM5.2, CK5/6, calretinin, HBME-1, thrombomodulin, WT-1, EMA, CEA, CD15, B72.3, BG8 and TTF-1). Regression and classification tree-based methods were applied to select the best combination of markers. The modelling procedures used employ successive, hierarchical predictions computed for individual cases to sort them into homogeneous classes.

RESULTS: Labelling for calretinin and lack of labelling for BG8 were sufficient for definite correlation with a diagnosis of malignant mesothelioma. CD15 provided further differentiating information in some cases.

CONCLUSION: A panel of three antibodies was sufficient in most cases to diagnose, or to exclude, epithelial mesothelioma. Calretinin exhibits the strongest correlative power of the antibodies tested.

KEYWORDS: Pleural malignant mesothelioma, antibody, calretinin, panel, metastatic adenocarcinoma, tree regression analysis
INTRODUCTION

The biopsy diagnosis of pleural malignant mesothelioma (MM) can be problematic and requires the use of ancillary techniques more frequently than most other epithelioid tumors. In most laboratories, immunohistochemistry is today the mainstay for the pathological diagnosis of MM. 1-4 Because no single antibody has been identified with 100% sensitivity and 100% specificity for a diagnosis of MM, panels of antibodies that include both positive and negative markers are routinely employed: yet even the most recent and comprehensive of the published studies focus on the sensitivity and specificity of single antibodies investigated independently of the others 4, and numerous recent reviews suggesting various panels of antibodies 5-13 are based on this type of analysis.5, 6, 14 A recent attempt at meta-analysis has been made to provide guidance,9 but because of heterogeneity in the raw data in the individual studies on which the meta-analysis was carried out, the validity of this approach is limited. The same principle applies to the website STATdxPathIq (formerly known as ‘Immunoquery’),15 which provides sensitivities and specificities of antibodies, based on published studies, and suggests IHC panels for differential diagnosis based on those data. Only a handful of studies have attempted to use more advanced statistical methods, including logistic regression 1, 2, 16, but the only study attempting to include stepwise logistic regression employed pleural effusion fluids and used a panel of antibodies that included predominantly positive carcinoma markers, relying on lack of labelling for a diagnosis of mesothelioma. 1

In this current study, a panel of mesothelial-related antibodies and carcinoma-related antibodies, as well as two general epithelial antibodies was used (Table 1) and a multivariate
statistical analysis—which involves observation and analysis of more than one variable at a time while taking simultaneously into account the effects of all variates on the endpoint of interest—was carried out. Therefore, the present study investigated a comprehensive antibody panel jointly. Constructing regression trees may be seen as a type of variate selection procedure. The aim is to differentiate reliably between pleural epithelial MM and secondary adenocarcinoma affecting the pleura. The objective was to ascertain which specific minimal set of markers proved most reliable in making the distinction between these two malignancies and correlating with a diagnosis of MM. This approach allows a rational decision for a minimal set of antibodies as primary line of investigation, which is known to maximize the diagnostic yield. This may reduce the cost of the investigation and the time required for the pathologist to assess the sections—an inconvenient but important rational consideration in a climate of ever increasing pressures associated with time constraints and cost savings. In other words, this approach aims to maximize the number of definitive diagnoses that can be made based on a specific limited panel of antibodies. If no definite diagnosis is reached after using the specified antibody panel, further immunohistochemical studies or other ancillary studies including electron microscopy may then be carried out as necessary.

MATERIALS AND METHODS

Histological Samples

A data set of 200 consecutive cases of pleural malignancy, for which there were comprehensive immunohistochemical data, was investigated. The cases were sourced from routine surgical specimens submitted to the Department of Anatomical Pathology at Flinders Medical Centre, South Australia, and referral cases to one of the authors (DWH). There were
173 cases of definite epithelial MM and 27 cases of secondary adenocarcinoma. The bias of mesothelioma cases was due to the nature of our referral practice. This was taken into account in the statistical modelling approach (*vide infra*).

The diagnosis of epithelial MM was defined according to currently accepted morphological criteria, 11, 17-22 including the radiological findings, histological appearances, mucin histochemistry, immunohistochemistry and, in many instances, electron microscopy—including ultrastructural examination of deparaffinized tissue when available and when there remained any doubt about a definite diagnosis. EM was used in particular when the histological findings were atypical and/or whenever immunohistochemistry yielded one major discordant result (e.g., positive labelling with B72.3 in a suspected MM) or two or more minor discordancies (e.g., equivocal staining for B72.3 and CD15 in a suspected MM). All tissues had been fixed in 10% buffered formalin and had undergone standard processing and embedding in paraffin wax. Sections were cut 4µm thick, deparaffinized and rehydrated, if blocks had been received. For some referral cases, only unstained spare slides were received (instead of a paraffin block) and immunohistochemical studies were performed on those cases as necessary. Some of the referral cases were sent including the immunohistochemical preparations that were done in the source laboratory, but with insufficient spare slides and no paraffin blocks, such that only limited immunohistochemical studies could be carried out. This accounts for the missing data (*vide infra*). The panel of antibodies used included the low molecular weight cytokeratin antibody CAM5.2 as a general epithelial marker, antibodies against epithelial membrane antigen (EMA), five mesothelial cell markers and five carcinoma-related markers (Table 1). Antigen retrieval was individualized for each antibody, and incubation with all primary antibodies was overnight (see Table 1 for details of dilutions of antibodies, source and antigen retrieval used). For the
first 69 cases the streptavidin-biotin-peroxidase complex method was used (Ultra Streptavidin Detection System, Signet Laboratories) as a detection system, whilst for the remaining cases the Dako Cytomation EnVision + Dual Link System (DAKO) was used. This system was introduced into the Department because it reduces problems with background staining relating to endogenous biotin, particularly in small biopsies. The primary antibodies were still used at the same dilution as before and the retrieval systems remained unchanged. All sections were assessed independently by two investigators (SK and DWH) and immunohistochemical labelling was scored on a 3-point ordinal scale: positive labeling = 1, equivocal labeling = 0.5, no labeling = 0. Equivocal labelling was defined as positive labelling in less than 2% of tumour cells or when it was uncertain whether trace staining represented genuine labelling or simply high background staining.

Statistical methods

Regression and classification tree-based modelling was applied in the statistical data analysis. This is an explanatory technique for uncovering structure in data, useful for both diagnostic and prognostic regression problems. 23, 24 In growing a regression tree, the process is continued until the ‘terminal’ node is homogeneous enough or it contains too few cases (≤ 5 by default). The model is fitted using a binary recursive partitioning algorithm, whereby the data are successively split along coordinate axes of the predictor variates (presence of labelling with different antibodies) so that, at any node, the split that maximally distinguishes the response variate (diagnosis of MM) in the left and the right branches is selected. The decision to branch was based on whether a so-called deviance statistically exceeded a cut-off value, which was chosen to optimize the modelling.
Deviance measures the fit of a statistical model to the data when the parameter estimation is likelihood-based—i.e., it is a measure of node heterogeneity. The fit is inferred to be good when the deviance equals its expected value which is the number of degrees of freedom (df), that is the number of cases available for estimating the model minus the number of model parameters. The deviance can also be employed to quantitate the significance of individual antibody markers. This is achieved by computing twice the log-likelihood of the ratio of the best model to that of the current model.

The modelling approach followed was first to fit an overly complex tree, and then ‘pruning’ the tree down to a suitable size. The tree-construction process has to be seen as a hierarchical refinement of probability modeling. If the sample numbers are sufficiently large, the study yields unbiased estimates of the disease classification probabilities (and misclassification rates). Any split which did not improve the model fit by a default factor of the complexity parameter was pruned off by (10-fold) cross-validations employed to ensure the numerical stability of the terminal nodes.

An attractive property of the method is that at each node of a classification tree there is a probability distribution over the classes. The prediction probability available for each terminal node is constant, but remains dependent on the structure of the tree (viz. its depth). It follows that the interpretation of this probability may not be exactly the same as the one provided by the logistic (discriminant analysis) risk model, which was also fitted to the data. The action taken for handling missing values was to retain cases with partially unavailable marker values (because this most closely represents the clinical situation where, for example in referral practice, complete sets of immunohistochemical studies are not always available). The method used surrogate rules if the splitting variate was unavailable. The strategy was to pass a case down the tree as far as it will go. If it reached a terminal node, a predicted probability
of 'caseness' (mesothelioma or adenocarcinoma) was computed for it. Otherwise the
distribution at the node reached was employed to predict the outcome. This is perhaps the
single most useful feature of the applied modeling approach, considering there were generally
some (2 to 6%) missing values on the antibody markers. For some markers the number of
missing observations was very large (over 70% for TTF-1) so as to render them practically
useless for the tree analysis.

In the application, the recursive partitioning method was used as implemented in the rpart
routine (http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cmf) of the S-
PLUS system.25

We also evaluated the statistical efficacy of the investigated panel of markers. For the
malignant mesothelioma markers, we report sensitivity (Se) and specificity (Sp) rates,
whereas for the anti-markers or adenocarcinoma markers we use their complement values
(i.e., 1-Se and 1-Sp). The cutpoint for a classification was chosen based on the tree regression
for the individual markers, with scoring: 1 = positive, 0.5 = equivocal, 0 = negative. The
binary cutpoint was decided by the algorithm, that is whether to align the indeterminate score
0.5 with either 0 or 1.

In clinical practice, however, clear positive staining is required for a result to be regarded as
positive. In situations where this decision resulted in a discrepancy with the algorithmic split,
the computations were also done with equivocal staining scored as negative. The clinical
approach to diagnosis has the effect of lowering the Se- and raising the Sp-values.

RESULTS
Sensitivity and Specificity

Table 2 presents in the conventional way the results of the sensitivities and specificities of the antibodies used. The highest sensitivities of mesothelioma markers were observed for calretinin (Se = 98%), while in the case of adenocarcinoma markers the highest sensitivity was for CEA (1-Se = 100%, indicating that none of the mesotheliomas showed positive labeling for this antigen) and B72.3 (1-Se = 98%). The scoring of equivocal labelling (0.5) as positive altered these results only marginally.

Logistic Regression Analysis

Table 3 presents the results of fitting a logistic model with mesothelioma as the clinical endpoint. The ‘intercept’ is the log odds for the subpopulation with all the other terms set equal zero. The null model with only the intercept term had a deviance of 138 on 179 (= 200 - 20 - 1) df; 20 cases were deleted due to missing values. When this deviance was subtracted from the deviance of the updated model with four additional significant parameters it yielded a residual deviance of 20 on 175 (= 179 - 4) df. This indicates a very good model fit as the value of the deviance statistic is much less than its expected value, viz. the df. The identified set included the positive marker calretinin plus three carcinoma-related markers which generally fail to label MMs (with negative coefficients, indicating that lack of labeling would support a diagnosis of mesothelioma) for epithelial MM. The change in deviance, that is 138 - 20 = 118 with four df, proves that the four predictors jointly form a statistically significant panel of antibodies. The sensitivity of the logistic model to predict MM was almost 99%, while the specificity was 88%.

Regression and Classification Tree Analyses
Classification tree is the most common use of tree-based methods whose endpoint is a factor giving the diagnostic categorisation of the patients. It is also possible to construct regression trees in which terminal node gives the predicted (numerical) value. The ideas for classification and regression tree stucture are quite similar, but the terminology differs. The justification for regression analysis is that it serves as the basis for the classification tree-construction procedure, so we consider it first. Table 4 presents a regression tree-based analysis of antibody combinations for predicting the clinical outcome of the 200 patients (no cases were excluded in the final model because of no missing values in these four markers).

At the root (i.e. top node of the tree) the predicted probability of MM (coded as a binary outcome variable) was 86.5%. This percentage represents the actual relative frequency of mesothelioma cases in the combined study series. Starting from this empirical setting, the ranking of the markers in the order of their improvement of the deviance in separate primary splits was the following: 1) calretinin, 2) CD15, 3) BG8, 4) B73.2, and 5) HBME-1. The data partition ended up in four terminal nodes defined by the three markers calretinin, CD15 and HBME-1. To be 99% sure of a mesothelioma endpoint, the model established that the following three conditions were sufficient (splitwise probability in parenthesis): calretinin > 0 (97%), CD15 = 0 (99%). In the leftmost branch, a case correlated almost certainly with a 'not MM' diagnosis, based on the indications of calretinin = 0 and HBME-1 < 1. The prediction based on regression tree was almost the same as with logistic regression. These results can be explained by the fact that calretinin was so important factor that it dominated the role of the other antigens. In a less clear circumstance, the different approaches would probably give a more discrepant outcome.
Table 5 presents the corresponding classification tree-based analysis for a differential diagnosis. With an equal prior probability distribution (0.5, 0.5) for MM and adenocarcinoma cases (defined as categorical outcome variable), the branching terminated in a structure depicted in Figure 1 with four terminal nodes. In the selection of three markers, the predominant role of calretinin remained the same, but now CD15 complemented BG8, both of which are adenocarcinoma-related markers ('negative' MM markers). The model predicted probability of MM diagnosis based on the three antibodies was estimated to be 100% (correct). We note parenthetically that using the priors proportional to the empirical class counts produced a classification tree very similar to the regression tree of Table 3. Therefore, the bias towards MM cases in our collection of cases did not affect the validity of the modeling process.

DISCUSSION

Comparison of the findings in this study with those in the published literature

Although many studies have evaluated the differential diagnostic value of antibody sets in the differential diagnosis of epithelial MM versus adenocarcinoma, no conclusive shortlist of antibodies has been identified to date, and there is no consensus on the optimal antibody panel. There are numerous problems with the published series available: in the directly comparable series, the case numbers have not always been large. For example, one study that concluded that the sensitivity of the mesothelioma-related antibody HBME-1 for malignant mesothelioma was 100% 26 based its findings on a series of only 17 cases. Other studies have attempted meta-analysis of all published studies, but this approach suffers from the inability to control the quality of the primary data. Such difficulties include the diversity of material included, with cytology specimens, autopsy material and surgically removed tissue being compared indiscriminately. Furthermore, different studies have used different sources
and dilutions of antibodies, different antigen retrieval techniques and different secondary antibodies. Different researches have followed different criteria for assessing positive labeling, and so forth. Some of these difficulties are particularly apparent with two of the antibodies that we found to be valuable in the diagnosis of MM, namely HBME-1 and calretinin.

*Calretinin:* Our results confirm some of the observations made by other investigators. A considerable number of studies over recent years have identified calretinin as a particularly useful marker, with a number of studies identifying 100% of MM cases as positive for calretinin 13, 27-30. This may be in part due the fact that we restricted our carcinoma-group to adenocarcinomas, and we applied strict criteria for assessing positive labelling, in that positive labelling of nuclei was required for a result to be designated as positive, whereas some previous publications have accepted cytoplasmic staining only, resulting in higher numbers of lung adenocarcinomas labelling with calretinin. Like us, other investigators have found high specificity for mesotheliomas, if nuclear labelling was required for a positive result, irrespective of cytoplasmic staining.7 It is worth noting that the same clone of antibody was used for all of our cases (Table 1). Previously, one group31 considered calretinin to be "useless" when a Chemicon guinea pig antibody was used but conceded that, when a different antibody was used, it showed "the highest sensitivity for mesothelial cells."31

The high sensitivity and specificity we found with calretinin may of course in part related to the selection of our cases: some tumours known to show positive labeling for calretinin, such as squamous cell carcinoma of lung origin and which may label for calretinin in up to 40% of cases, 32, 33 were not included in our study which assessed only the value of antibody panels for the distinction between mesothelioma and adenocarcinomas metastatic to the pleura.
The diagnostic utility of calretinin, amongst other markers, was reviewed in two recent publications 34, 35 which were based essentially on the same published reports and data, and although both studies drew slightly different conclusions in their assessment of diagnostic usefulness of this antibody, they both regarded calretinin as a 'useful' positive mesothelioma marker. Interestingly, one of those studies then concluded by recommending panels of antibodies 'based on their sensitivities and specificities' but, like so many other studies, not taking into account the complex relations between antibody reactions. In contrast to those papers, which attempted to analyze data collected over many years, from many different laboratories and assessed and scored by many different investigators, our study enrolled 173 consecutive cases of epithelial MM and applied a recursive partitioning algorithm to select a probabilistically-founded array of reliable correlative markers for the type of malignancy present. There is only one other study that attempted standard multiple logistic regression analysis, which was performed on various indicator combinations:2. That analysis pinpointed calretinin as an important marker, and can thus be compared to a selected list of markers in our corresponding logistic analysis. It is reassuring that our more sophisticated statistical approach confirms the role of calretinin as a useful positive marker for the differential diagnosis between epithelial mesothelioma and adenocarcinoma. In fact, in our series calretinin emerged as the premier correlate, and this had the effect of rendering marginal the supplementary information provided by the two next-best selected markers BG8 and CD15.

BG8 has been found useful in the distinction of adenocarcinoma from epithelioid MM.5, 9, 36 In a study investigating 12 antibodies and using regression analysis, BG8 was found to be one of the three most useful antibodies.7 However, in contrast to our study cytology samples were used, and the statistical approach used also differed from the approach used by us, discussed in detail below, with regards to calretinin. Nonetheless, it is reassuring that others have also proven the value of this antibody.
Apart from the majority of adenocarcinomas, BG8 labels 80% of squamous cell carcinomas. It does not label renal cell carcinomas, which were not included in our study, and this may in part account for the high sensitivity and specificity we found with this antibody.

CD15 has been established for over 20 years as a positive adenocarcinoma marker for the distinction from mesothelioma, despite some authors reporting positivity in up to 32% of malignant mesotheliomas. Like us, others have found that MM is only rarely positive for CD15. Again, the relative high sensitivity and specificity for this antibody may be in part related to the fact that we excluded squamous cell carcinomas, which are generally not positive. The majority of recent studies have found CD15 to be undetectable in most or all mesotheliomas investigated, while it was found to be expressed in all or most adenocarcinomas, particularly those of pulmonary origin. It has been noted that CD15 expression is often focal and that false-negative reactions can be related to small biopsies, and the high sensitivity and specificity seen in our study may be repeated to the fact that our study consisted predominantly of larger video-assisted thoracoscopy (VAT) biopsies.

Interestingly, the one study that used logistic regression on a panel that included CD15 found other antibodies more suitable for the diagnosis of adenocarcinoma, despite high sensitivity and specificity of CD15. As mentioned above, that study used cytology samples and used an antibody panel, which was heavily weighted towards carcinoma-related markers with only one mesothelial-related antibody (EMA with a membranous pattern of staining). In addition, the statistical approach used also differs from the approach used by ourselves: the authors applied a stepwise discriminant logistic regression analysis to select the 'parameters' of the model function for a differential diagnosis between mesothelioma and adenocarcinoma cases,
which is essentially a backwards-selection approach. In contrast, our study used a forward variable selection approach and based the intercept on the distribution of diagnoses in our collection of cases.

**Particular versus General Case Series**

Our alternative classification or regression tree-based analysis began at the root of the tree with an *a priori* probability of a case being an epithelial mesothelioma case at 0.865. From this starting point, the graph set forth a hierarchical construct for the correlative values of the antibody probe markers calretinin, CD15, and BG8 or HBME-1. However, the prevalence of mesotheliomas in our series does not represent the relative frequency of mesotheliomas *versus* adenocarcinomas in general clinical practice (about seven secondary adenocarcinomas for every mesothelioma). 46 On the other hand, not all cases of metastatic carcinoma come to biopsy, due to the known clinical history. We pooled all the adenocarcinoma cases, regardless of primary tumour site, which for obvious reasons does not provide a homogenous immunohistochemical profile, but the scenario of a 'malignant pleural effusion secondary to metastatic carcinoma of unknown primary site' represents a common problem. This approach to the unknown primary tumour resembles the true clinical dilemma. One of the strengths of our study is that we used only sections of the actual pleural deposits of secondary carcinomas, as opposed to performing immunohistochemical studies on the sections of primary carcinoma which has been done in some other studies, 47, 48 because there is always a possibility that the immunohistochemical profile for the pleural deposits is a little different from that of the primary tumour.

Therefore, one may ponder whether the specified assumption of prior probabilities of 'caseness' affects the tree analysis qualitatively, or how much it changes the quantitative
values of the nodal markers. The rpart model allows the specification of optional prior parameters. For the case of equal priors (i.e., the probability of being a mesothelioma or an adenocarcinoma is 0.5), produce a different result, both in terms of the hierarchical ordering of the markers and the magnitude of their prediction probabilities. Basically, the statistical inference question is whether the study aims to address a particularistic (descriptive) problem or a general (causal) one. In other words, any such case series that come from hospital-based and referral populations are somewhat specific place-specific and time-specific. However, this is no different to the variability between sensitivities and specificities between the same antibody in the hands of different investigators.

Also, because the settings in particular places (say, Adelaide, Australia v. Houston, Texas) are never absolutely the same, it is not surprising that different studies, using representative samples from the study populations, can produce discrepant results and inferences. Therefore, in investigating the problem of the role of the various antibodies in the diagnosis of mesothelioma, the relative sizes of the compared series is a matter of study design, which the investigator can choose. Assuming an equal (or ‘ignorance’) distribution is theoretically appropriate when little is known beforehand about the tumour in question.

**Stability of the Statistical Model**

Both the logistic regression and the tree-based methods identified the same three antibodies (calretinin, CD15 and BG8) that can be used to diagnose or exclude epithelial MM. However, the constructed tree can extract complementary information over the traditional regression analysis. This is conveyed, especially, by the hierarchical ordering of the included variates that lends itself to evaluation of the marker's predictive importance. The basic rule is that the closer to the root node a factor appears, the more important is its effect on the outcome.
Moreover, the dichotomous representation of markers allows convenient characterizations of individuals falling below or above a certain cutpoint. On the other hand, the relation between a marker and the outcome may not be natural in the sense that the cutpoint divides the group essentially into two homogeneous subgroups regarding the marker's influence on the outcome. In these circumstances, deciding on the existence of a suitable cutpoint is problematic. The recursive partition algorithm is a data-driven procedure and as such it behaves, by and large, as a black-box system concerning the choice of an adequate cutpoint. In the present study, the antibody labelling used only three score values. Therefore, the constructed tree is likely to be stable.

Generally speaking, the process of model selection may be viewed as a pattern-recognition process. To assert that a model fits the data well using some appropriate criterion in the ordinary sense, means that the data appear consistent with the pattern predicted by the model. Conversely, to assert that the model fit is poor means that the data appear to deviate appreciably from the model pattern. Model correctness cannot be inferred from the fact that the fit of the model to the data is good, since there are alternative models that may also provide a good fit. A good fit is, however, a necessary condition for inference.

In the context of predictive modelling, where the objective is interpretation, given specific states of knowledge, the function of ‘automatic’ procedures for the selection of variates into a model (such as stepwise forward selection in logistic regression) seems limited. In principle, all simple models adequately fitting the original data set (statistical validation) should be listed, and any choice between them should be made on how successfully a model performs on a new case series (clinical validation). Issues of interaction between variates are handled implicitly; there is no need to enter product terms in the model formula. The questions are
reduced to which variates to divide on, and how to achieve the split. The justification for the tree-based methodology is to view the tree as providing a probability model. The decision on a split is made, based on a change in a deviance measure for the tree under a likelihood function that is conditioned on a fixed set of observed random variates. Note that once fixed by observation, a random variate is in no sense variable, so it might better be called a correlative or predictive marker. The unknown prediction probabilities are estimated from the proportions in the split node. The tree construction process chooses the split according to the maximum reduction in the deviance measure. When the objective is prediction, classification error rate is the appropriate criterion for judging any particular model form.

**Problem of overfitting predictive models**

There are deficiencies in the standard modelling methods. It is well known that analyses that are not pre-specified but are data-dependent and liable to lead to overoptimistic conclusions. Many applications involve a large number of variates to be modelled using a relatively small patient sample. In the case of very expressive models such as regression trees, there is the danger that the models come up with chance idiosyncrasies of the particular data, which are not true in general. It is also important that the minimum size of the terminal nodes is large enough. Problems of overfitting and of identifying important markers are exacerbated in predictive modelling, because the accuracy of a model is more a function of the number of events than of the total sample size. Experience indicates that a complex model is more likely to give overoptimistic prediction when extensive variate selection has been done.

A related problem of variable selection is multicollinearity. We encountered a high correlation within clusters of positive and negative mesothelioma antigens and vis-à-vis the clusters.
Thus it was difficult to disentangle their individual effects and impossible to identify a unique solution for the regression tree. This means that correlated markers, epitomised by $r(CEA, WT-1) = -0.9$, may be essentially measuring the same underlying pathology or construct. In statistical and predictive terms, they both convey essentially the same information, although this may obviously not be true in a biological sense.

*Comparison to other statistical models*

Our type of tree-modeling logistic regression analysis is different from the standard multiple logistic regression analysis by Carella et al.2, which was performed on various indicator combiners. It is nonetheless reassuring that their analysis also pinpointed calretinin as an important marker, and their approach can be compared to a selected list of markers in our corresponding logistic analysis. As mentioned above, Dejmek and Hjerpe 1 applied a backwards selection based regression analysis approach, whilst our tree-construction process is essentially a forward variable selection. We favour the forward selection approach on intuitive grounds. Apart from the direction of selection, the adjustment of the intercept parameter (simply, the average proportion of mesothelioma) was adjusted in a manner to represent the proportion in the general population, instead of that in the study series.

An attractive property of our statistical approach is that at each node of a classification tree there is a probability distribution over the classes. The prediction probability available for each terminal node is constant, but remains dependent on the structure of the tree (related to its depth). It follows that the interpretation of this probability may not be exactly the same as the one provided by the logistic regression analysis risk model, which was also fitted to the data.
The action taken for handling missing values was to retain cases with partially unavailable marker values. The method used surrogate rules if the splitting variate was unavailable. The strategy was to pass a case down the tree as far as it will go. If it reached a terminal node, a predicted probability of caseness was computed for it.

CONCLUSIONS

The factors underpinning the selection of antibodies in large panels of antibodies for the distinction between MM of epithelial type and metastatic adenocarcinoma are often based on individual antibody's characteristics rather than the behavior of the whole panel. This study provides a novel and unique approach to this situation, because advanced tree-based partitioning methods are decidedly different from customary regression methods for predicting class membership on a binary response. Unlike any other statistical approach applied to this problem previously, our approach employs hierarchica modeling, with successive predictions being applied to particular cases, to sort the cases into homogeneous classes. Traditional methods use simultaneous techniques to make one and only one prediction for each and every case.

Our findings, based on a large prospective study encompassing 200 consecutive pleural biopsies, indicate that a panel consisting of three antibodies, namely calretinin, BG8, and CD15 are jointly sufficient to accurately diagnose or exclude a case as a primary epithelial MM in the majority of cases. This correlates quite closely with the recommendation from the International Mesothelioma Panel that at least two mesothelial cell markers and two carcinoma-related markers be used for the immunohistochemical investigation of suspected epithelial mesothelioma. However, we emphasise that this study represents a retrospective
statistical correlation using cases with a known diagnosis of mesothelioma. We will now use the three markers (calretinin, CD15, BG8) as a prospective first-line approach to mesothelioma diagnosis and then ascertain if the use of additional markers influences the confidence index for a diagnosis of mesothelioma. However, this approach is only suitable for a differential diagnosis of epithelial mesothelioma versus adenocarcinoma. In clinical practice, often a much wider differential diagnosis including epithelioid hemanagioendothelioma and biphasic synovial sarcoma needs to be considered. We also emphasise that the findings in this study are valid for our department, but because of the differences in methodology for immunohistochemistry (for example variable tissue processing, antigen retrieval, dilutions of primary antibodies and different detection systems, even when using the same antibodies produced by the same manufacturer), other laboratories may record different findings. Ideally, each laboratory should establish their own optimal panel of markers for mesothelioma diagnosis. Our rational and systematic approach to mesothelioma diagnosis indicates that overly exhaustive panels of antibodies may not improve the diagnostic confidence, and in our laboratory we will proceed with the prospective study of a limited panel of three markers, as indicated above. We believe that this approach deserves wider application for the immunohistochemical diagnosis of a variety of tumours.
REFERENCES

8. Ordonez NG: The immunohistochemical diagnosis of epithelial mesothelioma. [Review] [107 refs], Human Pathology 1999, 30:313-323
43. Wick MR, Loy T, Mills SE, al e: Malignant epithelioid pleural mesothelioma versus peripheral pulmonary adenocarcinoma: A histochemical, ultrastructural, and immunohistologic study of 103 cases, Human Pathology 1990, 21:759-766
44. Ordonez NG: Mesothelioma with rhabdoid features: an ultrastructural and immunohistochemical study of 10 cases, Mod Pathol 2006, 19:373-383
47. Ordonez NG: Value of the MOC-31 monoclonal antibody in differentiating epithelial pleural mesothelioma from lung adenocarcinoma, Human Pathology 1998, 29:166-169
53. Ordonez NG: Role of immunohistochemistry in distinguishing epithelial peritoneal mesotheliomas from peritoneal and ovarian serous carcinomas. [Review] [85 refs], American Journal of Surgical Pathology 1998, 22:1203-1214
60. Renshaw AA, Pinkus GS, Gorson JM: HBME-1 aids in distinguishing mesotheliomas and adenocarcinomas of the lung and breast., Mod Pathol 1995, 8:152A
63. Ordonez NG: Role of immunohistochemistry in differentiating epithelial mesothelioma from adenocarcinoma. Review and update. [Review] [144 refs], American Journal of Clinical Pathology 1999, 112:75-89
65. Szpak CA, Johnston WW, Roggli V, al e: The diagnostic distinction between malignant mesothelioma of the pleura and adenocarcinoma of the lung as defined by a monoclonal antibody (B72.3), American Journal of Pathology 1986, 122:252-260
Table 1. Details of antibodies used, including source, antigen retrieval and dilutions.

<table>
<thead>
<tr>
<th>Antibody</th>
<th>General Details</th>
<th>Pattern of labeling</th>
<th>Antigen retrieval</th>
<th>Dilution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Markers Positive in Malignant Mesothelioma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calretinin (Zymed)</td>
<td>Currently regarded by many as the most sensitive and specific marker for MM.2, 12, 13, 51</td>
<td>Nuclear (+/− cytoplasmic)</td>
<td>Trypsin</td>
<td>1:500</td>
</tr>
<tr>
<td>CK5/6 (Chemicon)</td>
<td>Positive in most epithelial MM (negative in most adenocarcinomas), but positive in ovarian serous carcinomas 52-55</td>
<td>Membrane labelling</td>
<td>Citric acid pH6</td>
<td>1:100</td>
</tr>
<tr>
<td>WT-1 (Dako)</td>
<td>A protein expressed by some fetal tissues and adult mesothelium. Reportedly good sensitivity and specificity for epithelial mesotheliomas;7, 40, 56-59 not usually reactive with renal cell carcinoma</td>
<td>Nuclear labelling</td>
<td>Citric acid pH6</td>
<td>1:100</td>
</tr>
<tr>
<td>Thrombomodulin (Dako)</td>
<td>Avoids missing an epithelioid hemangioendothelioma</td>
<td>Membrane labelling</td>
<td>Trypsin</td>
<td>1:400</td>
</tr>
<tr>
<td>HBME-1 (Dako)</td>
<td>Raised from human mesothelial cell line, exact antigen not known, appears to be associated with microvilli. Variably regarded 9, 12, 26, 57.</td>
<td>Membrane labelling</td>
<td>No retrieval</td>
<td>1:15,000</td>
</tr>
</tbody>
</table>

Please note: data sheet suggests 1:50-100 dilutions.
<table>
<thead>
<tr>
<th>Markers Positive in Carcinoma</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CEA (Zymed)</strong></td>
<td>Strong, diffuse, linear labelling supports diagnosis of malignancy</td>
</tr>
<tr>
<td><strong>CD15 (Leu-M1) (Dako)</strong></td>
<td>Well characterized useful for distinction from renal cell carcinoma (most are positive)</td>
</tr>
<tr>
<td><strong>B72.3 (kind gift of Dr Cant)</strong></td>
<td>Complex glycoprotein expressed in breast Ca, well established in literature but variable reports</td>
</tr>
<tr>
<td><strong>BG-8 (Signet)</strong></td>
<td>LewisY antigen, labels adenocarcinoma, 5, 7, 9, 36, 66 and 80% of squamous cell carcinomas, no labelling of renal cell carcinomas</td>
</tr>
<tr>
<td><strong>TTF-1 (Dako)</strong></td>
<td>Very specific for primary lung carcinoma (or carcinomas of thyroid follicular epithelium)</td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td></td>
</tr>
<tr>
<td><strong>EMA (clone E29)</strong></td>
<td>Membrane-bound glycosylated phosphoprotein anchored to the apical surface of many epithelia.</td>
</tr>
<tr>
<td><strong>CAM5.2 (Becton-Dickinson)</strong></td>
<td>General epithelial marker</td>
</tr>
</tbody>
</table>
Table 2. Individual sensitivity (Se) and specificity (Sp) of immunohistochemical markers for epithelial mesothelioma and adenocarcinoma, based on 173 cases of epithelial mesothelioma and 27 cases of adenocarcinoma. Discrepant percentages resulting from scoring equivocal marker staining as negative diagnosis are given in parenthesis.

<table>
<thead>
<tr>
<th>Epithelial marker</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Total number of tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Se, %</td>
<td>Sp, %</td>
<td></td>
</tr>
<tr>
<td>CAM 5.2</td>
<td>100.0</td>
<td>0.0</td>
<td>194</td>
</tr>
<tr>
<td><strong>Mesothelioma marker</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calretinin</td>
<td>98.2 (95.2)</td>
<td>81.5 (85.2)</td>
<td>194</td>
</tr>
<tr>
<td>CK 5.6</td>
<td>96.6</td>
<td>57.9</td>
<td>137</td>
</tr>
<tr>
<td>EMA</td>
<td>90.9 (86.0)</td>
<td>7.7 (15.4)</td>
<td>190</td>
</tr>
<tr>
<td>HBME-1</td>
<td>89.2 (86.7)</td>
<td>76.0 (80.0)</td>
<td>191</td>
</tr>
<tr>
<td>Thrombomodulin</td>
<td>89.6 (79.1)</td>
<td>56.0 (68.0)</td>
<td>188</td>
</tr>
<tr>
<td>WT-1</td>
<td>77.8</td>
<td>88.9</td>
<td>126</td>
</tr>
<tr>
<td><strong>Adenocarcinoma marker</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B72.3</td>
<td>98.2 (98.8)</td>
<td>54.2 (50.0)</td>
<td>187</td>
</tr>
<tr>
<td>BG8</td>
<td>83.2</td>
<td>88.5</td>
<td>193</td>
</tr>
<tr>
<td>CD15</td>
<td>68.2</td>
<td>73.1</td>
<td>196</td>
</tr>
<tr>
<td>CEA</td>
<td>100.0</td>
<td>63.0</td>
<td>193</td>
</tr>
<tr>
<td>Ber-EP4</td>
<td>82.4</td>
<td>83.3</td>
<td>57</td>
</tr>
<tr>
<td>TTF-1</td>
<td>92.9</td>
<td>52.9</td>
<td>59</td>
</tr>
</tbody>
</table>
Table 3. A logistic regression analysis of multiple antibodies for the diagnosis of epithelial malignant mesothelioma.

<table>
<thead>
<tr>
<th>Regression term</th>
<th>Estimate</th>
<th>Standard error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.02</td>
<td>1.21</td>
<td>1.62</td>
<td>0.11</td>
</tr>
<tr>
<td>Calretinin</td>
<td>4.53</td>
<td>1.63</td>
<td>2.77</td>
<td>0.006</td>
</tr>
<tr>
<td>BG.8</td>
<td>-4.47</td>
<td>1.99</td>
<td>-2.25</td>
<td>0.026</td>
</tr>
<tr>
<td>B72.3</td>
<td>-3.55</td>
<td>1.62</td>
<td>-2.19</td>
<td>0.017</td>
</tr>
<tr>
<td>CD.15</td>
<td>-4.05</td>
<td>1.68</td>
<td>-2.41</td>
<td>0.030</td>
</tr>
</tbody>
</table>
Table 4. A regression tree-based analysis of the deviance of multiple antibodies for the diagnosis of epithelial malignant mesothelioma. *denotes a terminal node.

<table>
<thead>
<tr>
<th>Criterion of partition for nodes of the regression tree</th>
<th>Number of cases</th>
<th>Deviance statistic within the node</th>
<th>Predicted probability1 of caseness, %</th>
<th>Terminal node *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top node of the tree</td>
<td>200</td>
<td>23.4</td>
<td>86.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Left branch</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calretin = 0</td>
<td>26</td>
<td>3.4</td>
<td>15.4</td>
<td></td>
</tr>
<tr>
<td>HBME-1 &lt; 1</td>
<td>19</td>
<td>0.9</td>
<td>5.3</td>
<td>*</td>
</tr>
<tr>
<td>HBME-1 = 1</td>
<td>7</td>
<td>1.7</td>
<td>1.7</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Right branch</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calretin &gt; 0</td>
<td>174</td>
<td>4.9</td>
<td>97.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Left branch</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD15 &gt; 0</td>
<td>9</td>
<td>2.0</td>
<td>88.7</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Right branch</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD15 = 0</td>
<td>165</td>
<td>2.0</td>
<td>98.8</td>
<td>*</td>
</tr>
</tbody>
</table>

1Probability at the top of the tree equals the relative frequency of mesothelioma cases in the series of patients with pleural malignancies.
Table 5. A classification tree-based analysis of the deviance of multiple antibodies for a diagnosis of epithelial mesothelioma. *denotes a terminal node.

<table>
<thead>
<tr>
<th>Order and criterion of node partition</th>
<th>Number of cases</th>
<th>Prediction misclassification rate, %</th>
<th>Probability of mesothelioma caseness, %</th>
<th>Terminal node</th>
</tr>
</thead>
<tbody>
<tr>
<td>0) Root node</td>
<td>200</td>
<td>50</td>
<td>50 / 50</td>
<td></td>
</tr>
<tr>
<td>1 Left branch) calretinin = 0</td>
<td>26</td>
<td>8.9</td>
<td>97.2 / 2.8</td>
<td>*</td>
</tr>
<tr>
<td>1 Right branch) calretinin &gt; 0</td>
<td>174</td>
<td>10.6</td>
<td>15.9 / 84.1</td>
<td></td>
</tr>
<tr>
<td>2 L) BG8 &gt; 0</td>
<td>32</td>
<td>48.8</td>
<td>54.3 / 45.7</td>
<td></td>
</tr>
<tr>
<td>3 L) CD15 &gt; 0</td>
<td>7</td>
<td>33.0</td>
<td>82.8 / 17.2</td>
<td>*</td>
</tr>
<tr>
<td>3 R) CD15 = 0</td>
<td>25</td>
<td>28.6</td>
<td>35.8 / 64.2</td>
<td>*</td>
</tr>
<tr>
<td>2 R) BG8 = 0</td>
<td>142</td>
<td>0</td>
<td>0 / 100</td>
<td>*</td>
</tr>
</tbody>
</table>

1) Probability means relative frequency (empirical or theoretical), as in "the clinically observed relative frequency of mesothelioma cases in our series of patients with pleural malignancies was 0.865" or "the model predicted probability of mesothelioma diagnosis based on the three antibodies was estimated to be 100% (correct), that is the misclassification rate (frequency in the particular patient population experience) was zero".
Figure 1. A decision tree (based on Table 5) for the immunohistochemical assessment of epithelial mesothelioma, assuming equal likelihood of the tumour representing mesothelioma or adenocarcinoma (0.5 likelihood of being a malignant mesothelioma at the first node). Each of the following nodes then provides the allocation ratio of cases as either tumour (adenocarcinoma / mesothelioma) and provides the likelihood of a lesion being a MM sbased on the immunohistochemical labelling. The inferior nodes (ellipses) reached by those initial distinctions are further analysed by left and right splits. The terminal nodes (rectangles) give the final differentiation between the cases. The decimal figure in each node is the predicted probability (ranging from 0 to 1) that an individual case represents an epithelial MM. For example, if a tumour was negative for calretinin at the first node, there was a 2.8% chance (probability of 0.028) of the tumour being finally diagnosed as epithelial MM.
Marker scoring: 1 = positive, 0.5 = equivocal, 0 = negative. The binary cutpoint was decided by the algorithm (i.e., whether to align the indeterminate score 0.5 with either 0 or 1).