ELSEVIER

Journal of Clinical Epidemiology 54 (2001) 603–609

Journal of Clinical Epidemiology

The use of classification and regression trees in clinical epidemiology

Roger J. Marshall*

Department of Community Health, University of Auckland, Private Bag, Auckland, New Zealand

Received 16 February 2000; received in revised form 2 October 2000; accepted 4 October 2000

Abstract

A critique is presented of the use of tree-based partitioning algorithms to formulate classification rules and identify subgroups from clinical and epidemiological data. It is argued that the methods have a number of limitations, despite their popularity and apparent closeness to clinical reasoning processes. The issue of redundancy in tree-derived decision rules is discussed. Simple rules may be unlikely to be "discovered" by tree growing. Subgroups identified by trees are often hard to interpret or believe and net effects are not assessed. These problems arise fundamentally because trees are hierarchical. Newer refinements of tree technology seem unlikely to be useful, wedded as they are to hierarchical structures.

Keywords: Classification rules; Tree-based algorithms; Clinical epidemiology

1. Introduction

To establish diagnostic and prognostic rules and to determine high-risk groups in clinical and epidemiological analysis, researchers often turn to tree-based methods (sometimes called recursive partitioning) in preference to more well-established regression modeling. Statistical trees can be "grown" using various easy-to-use software packages. The appeal of the approach seems to arise from perceived weak, complex and unrealistic linear modeling methods compared with the apparent simplicity of a tree. In clinical medicine, the methods appear to have more affinity with the way clinicians make decisions—they logically demarcate clusters of signs and symptoms and so on, rather than make distinctions based on arithmetic scores. This was noted by Koss and Feinstein [1], who seem to be the first to suggest using trees to consolidate similar patient groups.

Trees have now been developed as diagnostic and prognostic tools in a host of clinical situations—for example, to identify high risk for myocardial infarction [2–4], evaluating chest pain [5], dental caries [6], asthma [7], diabetes [8] and others. They have also been proposed to identify subgroups that are at different levels of risk in clinical and epidemiological investigations [9–12].

Trees were originally proposed as a way of "automatic interaction detection" (AID) [13,14]. It was envisaged that analysis of survey data by this method would enable homogeneous population subgroups to be revealed. At the time, the method was criticized by Einhorn [15] as one of a number of "alchemical" methods. Einhorn suggested that the method may "make sense out of noise" since no specific functional form is imposed. Doyle [16] also criticized its potential for spurious results, and also noted that *net* effects of factors are not determined, a point also made more recently by Segal [17].

Nevertheless, the method took hold and was bolstered by the work of Breiman et al. [18], which laid a theoretical basis for dealing with the problem of overfitting, that is, finding a large tree that closely fits the data at hand but not new data, by cost-complexity penalizing. Indeed, Brieman's work has been seen as one of the most important advances in statistics in the 1980s [19]. At the same time as these statistical advances were progressing, developments in tree-based methods were also taking place in the artificial intelligence community [20]. The literature is now vast and spans machine learning and statistical journals; the two domains bought nicely together in comparisons in the so-called Statlog project [21].

But there remain, in my view, problems with trees, especially in their application in medicine, which have not entirely gone away, despite Breiman et al.'s pioneering work and many advances since. In this article I will draw attention to some of these.

2. Notation, trees and Boolean algebra

For simplicity, I will consider only binary trees, that is, those with only two-way splits at each node. I will also focus mainly on the two-class case when Y is binary. It is ex-

* Corresponding author.

0895-4356/01/$ – see front matter
PII: S0895-4356(00)00344-9

tremely common in medical studies to usually classify individuals as “high” or “low” risk (e.g. [2–8]).

Trees are generated by repeatedly splitting a sample. Each split is made in terms of an *attribute* that is either possessed or it is not by each individual in the sample. For example, *age over 50* could be such an attribute and *obesity* another. I will use capital letters to represent attributes.

For two attributes, A and B, having both is represented by: *A and B*. Possession of either one is: *A or B*. Possession of either A and B, or C and D is: (*A and B*) or (*C and D*). In order to simplify the notation in Boolean expressions, the *and/or* connectives will be dropped. So, for instance, (*A and B) or (C and D)* will be written simply as (AB)(CD). Also $\bar{A}$ will denote *not A*. I will refer to A as an attribute that is *present*, while $\bar{A}$ is an absent attribute.

Each terminal node in a tree can be represented by the conjunction of the attributes along the pathway to the node. Consider, for example, Fig. 1, which is a classification tree for predicting high risk of diabetes (redrawn from Herman et al. [8]). There are seven terminal nodes and four are designated high risk (shown by the boxes). The rightmost terminal node is the pathway *A and O*, or simply AO. The pathways to the other three high-risk terminal nodes are $\bar{A}OS$, $A\bar{O}H$ and $A\bar{O}\bar{H}G$.

To be classified as high risk an individual must arrive at either of the four high-risk terminal nodes. Accordingly, the decision rule for high risk can be written as the union of the terminal nodes.

$$(AO)(\bar{A}OS)(A\bar{O}H)(A\bar{O}\bar{H}G) \tag{1}$$

In Fig. 1, and all other trees that are presented, I adopt the convention: go to the right when an attribute is present, go to the left when it is absent. Then the leftmost node has to be a combination of absent attributes and the rightmost a combination of present attributes. Intermediate terminal nodes are a mix of present and absent attributes.

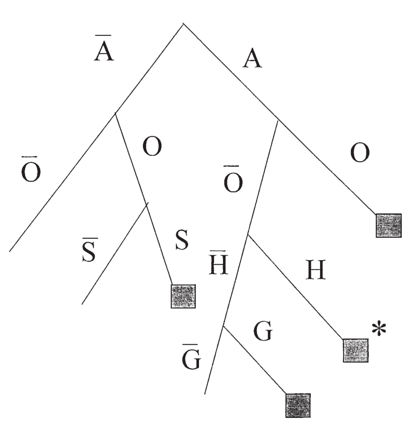

Fig. 1. Tree for classification to high risk of diabetes (redrawn from Fig. 2 in Herman et al. [8]). Box terminal nodes are high risk. Attributes are: *A*—age over 45; *O*—obese; *H*—hypertensive; *G*—glucose intolerance; *S*—sedentary. The node marked * is mentioned in the text.

3. Redundancy in decision rules

One of my criticisms of trees is that, in the process of taking the union of terminal nodes, the decision rule may simplify and clusters of attributes that define the terminal nodes, and that appear to be important, may be misleading.

Consider again the tree in Fig. 1. The box terminal nodes indicate high risk (prevalence of undiagnosed diabetes $\geqslant$5%) and the rule is given by expression (1). The rule indicates, for example, by the pathway $A\bar{O}H$ to the terminal node marked by an asterisk, that over 45-year-old hypertensives (*AH*) who are *not* obese ($\bar{O}$) qualify for the rule. However, this is a sufficient condition, it is not *necessary* to be nonobese to satisfy the decision rule (1); just being over 45 and hypertensive will do. This becomes evident by applying elementary, although tedious, Boolean algebra. Expression (1) reduces to

$$(AO)(OS)(AH)(AG) \tag{2}$$

No more simplification is possible. In this form, the expression shows that neither $\bar{O}$, $\bar{A}$ nor $\bar{H}$, which are in the original decision rule (1), are prerequisites of the rule at all. The residual combination (*AH*) indicates that being an over 45-year-old hypertensive is enough to qualify for the high-risk group; absence of obesity is immaterial.

Another example is the fairly complex tree in Fig. 2 for identification of children with high dental caries increment rate [6]. Writing the decision rule as a union of the pathways to the seven high-rate terminal nodes is somewhat messy, but here it is:

$$(\bar{A}B)(AC)(\bar{C}X)(\bar{C}\bar{E}Y)(\bar{C}\bar{E}Z)(\bar{C}\bar{E}\bar{G}W)(\bar{C}\bar{E}\bar{G}\bar{I}V) \tag{3}$$

where $X = ADE$, $Y = ADFH$, $Z = AD\bar{F}G$, $W = AD\bar{F}I$ and $V = (AD\bar{F}JK)$ (the meaning of each attribute is in the caption to Fig. 2). I have deliberately formed and substituted X, Y, Z, W and V, not just because the expression looks rather less formidable than it would otherwise, but also because the expression (3) actually reduces to

$$(\bar{A}B)(AC)(X)(Y)(Z)(W)(V)$$

In other words the attributes $\bar{C}$, $\bar{E}$, $\bar{I}$ and $\bar{G}$ are redundant in evaluating the decision rule. It is sufficient, considering the terminal node marked by an asterisk in Fig. 2 for example, that X is satisfied; it is not necessary that it be in conjunction with $\bar{C}$. The impression conveyed by the tree in Fig. 2 of the importance of $\bar{C}$, the branch from which five of the terminal nodes are grown, is misleading.

Again this example illustrates a difficulty with trees: the classification rule that is generated by a tree may simplify to make some attributes redundant in evaluating the rule. The impression may be created that certain combinations of attributes are important when they are not.

In the above, it is important to appreciate that the redundancy arises solely because the rule simplifies. It is not a data-driven redundancy, in the sense discussed by Quinlan [20]. He considers the effect, on prediction error, of remov-

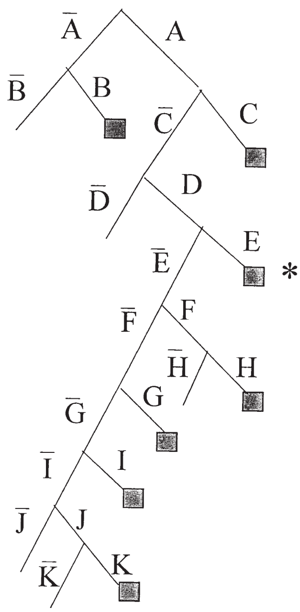

Fig. 2. Tree for classification to high risk of caries (redrawn from Fig. 4 in Stewart and Stamm [6]). Box terminal nodes are high risk. Attributes are: *A*—primary decayed, missing, or filled (dmfs) ⩾ 5; *B*—morphology score ⩾ 12; *C*—permanent dmfs ⩾ 1; *D*—fissured surface ⩾ 1; *E*—morphology score ⩾ 11; *F*—age ⩾ 6.7; *G*—age ⩾ 6.65; *H*—large decay ⩾ 15; *I*—minimal decay ⩾ 5; *J*—referral score ⩾ 2; *K*—primary fillings ⩾ 3. The node marked * is mentioned in the text.

ing attributes along the paths to terminal nodes. Removing attributes in this way will usually change the decision rule, though in some situations, as the two examples already discussed, it may not.

4. Regular decision rules

Usually a potential predictor of a binary, or continuous, outcome variable y is measured or recorded because there is, in advance, reason to suppose it is associated with the outcome. Also, the direction of the association can usually be assumed. It is questionable, scientifically, whether a factor should even be considered as a predictor unless there is reason to suppose it has an association and that we are reasonably confident of its direction. If not, we are truly "fishing." Therefore, I suggest, attributes can usually be labelled in advance as either *positive* or *negative* with respect to the outcome. Specifically, I will assume that attribute A is labelled as positive for the outcome when, a priori, it is believed that the expected value of the outcome y when A is present exceeds the expected value of it when A is absent.

I will therefore adopt the convention that all attributes A, B, C, . . . are positive and their complements $\bar{A}$, $\bar{B}$, $\bar{C}$, . . . are negative. For example, in the first example above, attributes A (age over 65), O (obesity), H (hypertensive), G (glucose intolerance), S (sedentary) are all positive (i.e., adverse with respect to diabetes). The rule in expression (2) is a combination of only positive attributes, even though expression (1), from which it is derived, is written in terms of both positive and negative attributes. Rules of the form of expression (2), that is, combinations of only positive attributes, have been termed "regular" and are the basis for one method to develop nonhierarchical classification rules [22–24]. The regular rule in expression (2) is easy to interpret because there are no combinations that conflict with prior labelling of attributes as being positive.

For example, if $(AH\bar{O})$, that is, absence of obesity in conjunction with age over 50 and hypertension, which is in the original rule (1), had remained in the rule after simplification, it would require an explanation. It would be reasonable to ask whether being hypertensive and over 50 really is a high-risk combination *only in the nonobese*. One might realistically expect it to also raise the risk in obese as well. In other words, if $AH\bar{O}$ were in the high-risk rule, you might expect AHO to also be in the rule. But, once AHO is added to the rule, obesity becomes irrelevant since $(AH\bar{O})$ *or* $(AHO) = AH$.

This argument was originally used as a justification for focussing on regular decision rules [22]. By the very nature of tree growing, in which you either go left or right at each node, a classification rule derived from a tree will not usually be regular; combinations of positive and negative attributes will usually remain, as in the second dental caries example where, despite $\bar{C}$, $\bar{E}$, $\bar{I}$ and $\bar{G}$ dropping out of the rule, $\bar{F}$ and F, and $\bar{A}$ and A remain in expression (3). That is, for example, in combination with certain attributes, being *over* 6.7 years old (F) raises the risk and, in combination certain other attributes, being *under* 6.7 years old ($\bar{F}$) raises the risk.

These "interactions" may, or may not, be real. Unfortunately, there is no way to test whether they are. They might simply be a consequence of the way trees are constructed by splitting subgroups, a process that inevitably leads to combinations of positive and negative attributes at terminal nodes.

In the diabetes example, it is somewhat remarkable that the decision rule (1) reduced to one that *is* regular. Usually tree-grown classification rules do not simplify to such an extent, and may not simplify at all.

Consider, for example, a tree for classification to high risk of hospitalization for asthma in Fig. 3 (redrawn from Lieu et al. [7]). With attributes as defined in the Fig. 3 caption, the classification rule to high risk is

$$(HC)(\bar{H}PU_9)(\bar{H}\bar{P}BU_2)$$

This expression does not simplify. It leaves the impression, in the element $\bar{H}PU_9$, for example, that having six or more physicians prescribe asthma medications (P) and nine or more urgent visits (U_9) is only important if the person has not been previously hospitalized ($\bar{H}$). This may possibly be true but since the initial split is on H, PU_9 can only be combined on the left with $\bar{H}$. Whether this is a combination that is clinically meaningful seems debatable; PU_9 may be just as important in previously hospitalized patients too.

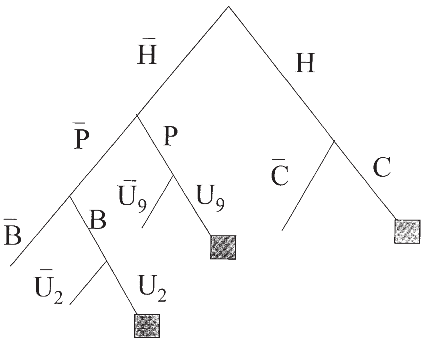

Fig. 3. Tree for classification to high risk of asthma (redrawn from Fig. 1 in Lieu et al. [7]). Box terminal nodes are high asthma risk classifications. Attributes are: *H*—hospitalized during prior 6 months; *C*—obtained two or more medication (cromolyn) units; U_9—made nine or more urgent clinic visits; *B*—obtained 16 or more units of beta-agonists; U_2—made two or more urgent clinic visits; *P*—six or more physicians prescribe asthma medication.

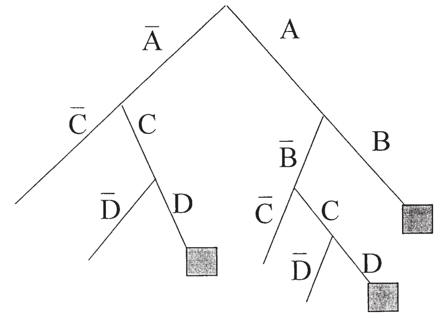

Fig. 4. Tree to represent the decision rule *(AB)(CD)*. Collecting together the boxed terminal nodes gives $(AB)(A\bar{B}CD)(\bar{A}CD) = (AB)(CD)$.

5. Simple rules and complex trees

Although tree grown rules may sometimes simplify to represent relatively simple rules, simple decision rules may actually require quite complex trees to represent them. This problem is recognized and has been called the replication problem [25].

Consider, for example, predicting low birth weight with the regular decision rule

$$(AB)(CD)$$

where positive attributes are: *A* is age under 18, *B* is black, *C* is cigarette smoker and *D* is drinker of alcohol. There are different ways that this rule can be represented by a tree. One is as in Fig. 4, with seven terminal modes, three of which are unioned to constitute *(AB)(CD)*, as indicated in the figure caption. Obviously, by symmetry considerations, other trees would give the same rule, although a simpler tree representations, that is, one with fewer than seven terminal nodes, is not possible, unless “Boolean splits” (see below) are allowed. Another example is the rule:

$$(ABC)(DEF)(GHK)$$

which apparently can only be represented by a tree with about 80 terminal nodes [26].

If these rules did provide excellent discrimination, it seems quite probable, and in the second case almost certain, that they would *not* be found by a tree search; one “wrong split” and the rules may not be discovered. Further, taking the first example, even if the generated tree structure was as in Fig. 4, the three terminal nodes, as indicated, would each have to be assigned to the same category to be amalgamated to create the rule.

Another issue that this raises is tree “complexity.” In the CART [18] approach to tree growing the complexity of a tree is represented by the number of its terminal nodes. As relatively complex trees may represent simple rules, the number of terminal nodes may not necessarily be an appropriate measure of the complexity of the decision rule, and penalizing by this measure may be unfair.

6. Tree models

In regression modeling, the model is intended as a plausible description of how an outcome may depend on a set of predictors. Although not usually presented as such, model fitting is, in effect, a procedure to search among all possible models of a certain form, for one that is best, based on some criterion (e.g., maximum likelihood or least squares).

Tree analysis is also a search procedure and the analogy with the set of all possible models is the set of all possible trees. One question that immediately arises is whether hierarchical disaggregation is the best way to conduct a search. As the number of all possible trees is generally extremely large, it is impractical to generate them all, or even a subset of them (if it were possible to sensibly define a subset), to see which is “best.” Tree growing has developed primarily because there seems to be no alternative. When it was first proposed [13], for instance, it was seen as second best to a search of “all possible combinations of characteristics.” However, there is no assurance that the generated tree is close to a global best, since it is derived from a series of local subsearches.

A second issue is whether the set of all possible trees is, anyway, too general, in the sense that it admits innumerable possible combinations of attributes. Just as a regression model imposes a certain structure, it seems preferable to consider a class of structured decision rules and to conduct a search among possible rules of that type. One class of structured rules, to which I have already alluded, is that of regular rules [22–24].

Imposing the regular structure, or some other, besides being scientifically sensible, reduces the size of the set of possible rules, so that a direct search becomes feasible rather having to search by disaggregation. Occasionally, as I have shown, tree-based decision rules may reduce to produce regular, or near-regular, rules, suggesting that a direct search of such rules would, from the outset, be a preferable strategy.

7. Regression trees

In the foregoing sections I have been concerned with tree-generated decision rules—sometimes called classification trees. So-called regression trees, on the other hand, often have a continuous outcome variable and are intended as an exploratory tool to discover homogeneous subgroups of the data. They are becoming popular in epidemiological investigations to discover high-risk groups [10–12] and in survival studies to determine poor prognosis groups [28–30].

Since the objective of a regression tree is formally to discover meaningful subgroups, rather than just work up a decision rule, being able to make sense and attach meaning to the terminal node (subgroup) combinations assumes prime importance.

With the exception of just two terminal nodes (the rightmost and leftmost, in the convention used here to draw trees), all others will be a mix of positive and negative attributes. As I have already intimated, these may be difficult to interpret or "believe." Also subgroups are, because of hierarchical splitting, mutually exclusive. By definition, therefore, the collection of attributes that define one subgroup are *not* possessed by any other. This feature is not necessarily desirable or useful. Often one may be interested in subgroups that *do* overlap. For example, a subgroup of under 20-year-old smokers may be at high risk of a low birth weight baby. Another high-risk subgroup may consist of black smokers. These two groups are not necessarily mutually exclusive, and their discovery would not be found in the terminal nodes of a tree, though they may in a direct search of regular Boolean structures [23].

8. Missed combinations

When tree growing was first proposed, as AID [13,14], it was recognized that interactions may fail to be "discovered" by the process of tree growing. Attributes which, by themselves, do not discriminate between classes and therefore do not form splits, may act in synergy with other attributes but such effects remain unseen. This issue still remains and, although programs may allow "Boolean splits" [18] or splits based on linear combinations of variables [18,31,32], which potentially deal with such circumstances, they seem to be seldom used in medicine and the resultant trees may be even more difficult to interpret.

9. Forest of trees

Cross-validation and bootstrap methods have been proposed to lend statistical validity to tree-based classification rules. Both methods rely on subsampling and growing new trees on subsamples. Trees grown on subsamples may differ, in topological structure, from those grown on the full sample, giving rise to a "forest of trees" [33].

The sensitivity of tree topology to data subsamples is both problematic and useful. On the one hand, it is useful since it allows unbiased estimation of error rates using cross-validation, which form the basis of the CART system of choosing a "right sized tree." In this procedure the estimated errors of trees grown on subsamples are pooled to estimate of error the tree grown on the whole sample.

On the other hand, sensitivity of tree topology presents an "embarrassment of riches" [33], that is, numerous different trees and the problem of which one to choose. Some elaborate methods have been proposed to make sense of the forest. For example, bootstrap aggregation ("bagging") [34,35] has been suggested: obtain a number of prediction rules from trees grown on random subsamples of the full sample and use each to classify a new individual. The predicted category that gets the most "votes" is the selected prediction. "Boosting" is another voting method [35]. It seems unlikely that clinicians would favor voting prediction rules. Alternatively, methods to select a single tree from a number of competitors have been advanced [33,36]. These require measures of "distance" and "similarity" between classifications made by the trees [33], or differences in tree topology [36]. The measures do not account for potential redundancy, or even equivalence, of decision rules from different trees.

10. Conclusion

Tree growing is one of a number of "data mining" techniques [27,37]. Some statisticians are scornful of these "data dredging" methods, but data exploration is inevitably an activity in which most responsible clinical and epidemiologic researchers engage. Tree growing is data dredging done in a systematic way and, in a sense, model fitting is also data dredging—dredging the parameter space to find where to maximize fit. It is not, in my view, dredging that is the problem, but that the tree paradigm imposes no sensible structure within which to dredge and that the dredging is done in a localized fashion.

For classification purposes in medical work, most reported studies show that there is really very little difference between the performance of logistic model methods and trees [3,5,9, 38–41]. The rules and subgroups that are derived from trees are lauded as being very easily described to clinicians since they logically demarcate clusters of symptoms, signs and other features. However, the clusters may not necessarily agree with established diagnostic combinations [42] and, as I have tried to show, clinicians may be misled into thinking the clusters have biologic meaning; they need to be regarded with skepticism. Feinstein [43] makes the similar point that tree-formed clusters may be composed of "heterogeneous constituents with no apparent biologic coherence."

Recent complex developments in tree technology (e.g., bagging [34], boosting [35], maximum likelihood [36]) seem unlikely to be able to deal with some of conceptual problems of trees, as the tree structure is itself the root (excuse the pun) of the most of the difficulties I have mentioned. To establish nonarithmetic diagnostic and prognostic combinations other non-tree search partitioning methods may be preferable.

One of these—search partition analysis (SPAN) [22–24]—has already been mentioned. Others are discussed in the machine learning literature [25,44]. In the medical context, Feinstein [43] has proposed "conjunctive consolidation" and "sequestered segmentation" to form homogeneous groups of patients. Both allow judgmental decisions about combining categories, recognizing the clinical and biologic context. In SPAN judgements must also be made about which attributes are positive, in the sense already discussed. SPAN generally performs as well as trees and logistic regression [22,24,45, see also http://www.auekland.ac.nz/mch/span/compspan.htm].

To do the algebraic simplifications that have been presented is somewhat tedious by hand, but can be achieved in the software used to implement the SPAN method, which is available from URL http://www.auckland.ac.nz/mch/span.htm.

References

[1] Koss N, Feinstein AR. Computer aided prognosis. II. Development of a prognostic algorithm. Arch Int Med 1971;127:448–59.

[2] Buntinx F, Truyen J, Embrechts P, Moreels G, Peeters R. Evaluating patients with chest pain using classification and regression trees. Fam Pract 1992;9:149–53.

[3] Tsien CL, Fraser HSF, Long WJ, Kennedy RL. Using classification tree and logistic regression method to diagnose myocardial infarction. In: Cesnik B, McCray AT, Scherrer JR, editors. Proceedings of the Ninth World Congress on Medical Informatics. IOS Press, Amsterdam 1998. p. 493–7.

[4] Goldman L, Weinberg M, Weisberg M, Olshen R, Cook E, Sargent RK, Lamas GA, Dennis C, Wilson C, Deckelbaum L, Feinberg H, Stiratelli R, and the Medical House Staffs at Yale-New Haven Hospital and Brigham and Women's Hospital. A computer derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. N Engl J Med 1982;307:588–96.

[5] Selker HP, Griffith JL, Sanjay P, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischaemia among emergency department patients. J Invest Med 1995;43:468–76.

[6] Stewart PW, Stamm JW. Classification tree prediction models for dental caries from clinical, microbiological and interview data. J Dent Res 1991;70:1239–51.

[7] Lieu CA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high rick children with asthma. Am J Respir Crit Care Med 1998;157:1173–80.

[8] Herman WH, Engelgau MM, Smith PJ, Aubert RE, Thompson TJ. A new and simple questionnaire to identify people at increased risk for undiagnosed diabetes. Diabetes Care 1995;18:382–7.

[9] Nelson LM, Bloch DA, Longstreth WT, Shi H. Recursive partitioning for identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. J Clin Epidmiol 1998;51:199–209.

[10] Zhang H, Holford T, Bracken MB. A tree-based method of analysis for prospective studies. Stat Med 1996;15:37–49.

[11] Carmelli D, Zhang H, Swan GE. Obesity and 33-year follow-up for coronary heart disease and cancer mortality. Epidemiology 1997;8:378–83.

[12] Zhang H, Bracken MB. Tree-based factor analysis of preterm delivery and small-for-gestational-age birth. Am J Epidemiol 1995;141:70–8.

[13] Morgan JN, Sonquist JA. Problems in the analysis of survey data and a proposal. J Am Stat Assoc 1963;58:415–34.

[14] Sonquist JA, Baker EL, Morgan JN. Searching for structure. Institute for Social Research, University of Michigan, 1971.

[15] Einhorn HJ. Alchemy in the behavioral sciences. Public Opinion Q 1972;36:367–78.

[16] Doyle P. The use of automatic interaction detector and similar search procedures. Oper Res Q 1973;24:465–7.

[17] Segal MR. Features of tree-structured survival analysis. Epidemiology 1997;8:344–6.

[18] Breiman L, Freidman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, CA: Wadsworth International, 1984.

[19] Tibshiriani R, LeBlanc M. A strategy for binary description and classification. J Graphic Comput Stat 1992;1:3–20.

[20] Quinlan JR. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1993.

[21] Michie D, Speigelhalter DJ, Taylor CC. Machine learning, neural nets and statistical classification. New York: Ellis Horwood, 1994.

[22] Marshall RJ. Partitioning methods for classification and decision making in medicine. Stat Med 1986;5:517–26.

[23] Marshall RJ. A program to implement a search method for identification of clinical subgroups. Stat Med 1995;14:2645–59.

[24] Marshall RJ. Classification to ordinal categories using a search partition methodology with an application in diabetes screening. Stat Med 1999;18:2723–36.

[25] Pagallo G, Haussler D. Boolean feature discovery and empirical learning. Machine Learning 1990;5:71–99.

[26] Wallace CS, Patrick JD. Coding decision trees. Machine Learning 1993;11:7–22.

[27] Gentleman R, Jorgensen M. Data mining. Chance 1998;11:34–9.

[28] Segal MR. Regression trees for censored data. Biometrics 1988;44:35–7.

[29] Marubini E, Morabito A, Valsecchi, MG. Prognostic factors and risk groups: some results given by using an algorithm suitable for censored survival data. Stat Med 1983;2:295–303.

[30] Ahn H, WY Loh. Tree-structured proportional hazards regression modeling. Biometrics 1995;50:471–85.

[31] Frank E, Wang Y, Inglis S, Holmes G, Witten IH. Using modal trees for classification. Machine Learning 1998;32:63–76.

[32] Brodley CE, Utgoff PE. Multivariate decision trees. Machine Learning 1995;19:45–77.

[33] Chipman HA, George EI, McCulloch RE. Making sense of a forest of trees. In: Weisberg S, editor. Proceedings of the 30th Symposium on the Interface. Interface Foundation of North America 1998 p84–92. Proceedings of the 30th Symposium on the Interface: 1998. p. 84–92.

[34] Breiman L. Bagging predictors. Machine Learning 1996;24:123–40.

[35] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. Machine Learning 1999;36:105–39.

[36] Shannon WD, Banks D. Combining classification trees using MLE. Stat Med 1999;18:727–40.

[37] MacKinnon MJ, Glick N. Data mining and knowledge discovery in databases—an overview. Aust N Z J Stat 1999;41:255–75.

[38] Cook EF, Goldman L. Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. J Chron Dis 1984;37:721–31.

[39] Werneck GL, de Carvalho DM, Barroso DE, Cook EF, Walker AM. Classification trees and logistic regression applied to prognostic studies: a comparison using meningococcal disease as an example. J Trop Pediatr 1999;45:248–51.

[40] Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the GUSTO database. Stat Med 1998;17:2501–08.

[41] Long WJ, Griffith JL, Selker HP, D'Agostino RB. A comparison of logistic regression to decision tree induction in a medical domain. Comp Biomed Res 1993;26:74–97.
[42] Chiogna M, Spiegelhalter DJ, Franklin RCG, Bull K. An empirical comparison of expert-derived and data-derived classification trees. Stat Med 1996;15:157–69.
[43] Feinstein AR. Multivariable analysis: an introduction. New Haven: Yale University Press, 1996.
[44] Rivest RL. Learning decision lists. Machine Learning 1988;2:229–46.
[45] Clarke DM, McKenzie DP, Marshall RJ, Smith GC. The construction of a brief case-finding instrument for depression in physically ill. Integr Psychiatry 1994;10:117–23.