Current Epidemiology Reports (2018) 5:221–229
https://doi.org/10.1007/s40471-018-0163-y

EPIDEMIOLOGIC METHODS (R MACLEHOSE, SECTION EDITOR)

Branching Out: Use of Decision Trees in Epidemiology

Julian Wolfson[1] · Ashwini Venkatasubramaniam[2]

Published online: 23 July 2018

Abstract

Purpose of Review Decision trees are a well-established tool for statistical modeling and machine learning, but they are not widely used in the epidemiological literature. In this review, we introduce the reader to the basic concept of the decision tree and describe three distinct ways that they can be used: for explanatory modeling, outcome prediction, and subgroup identification.

Recent Findings We discuss varieties and generalizations of decision trees that are best-suited for analyzing epidemiological data and introduce some visualizations which can help researchers interpret decision tree outputs. Throughout, we provide diverse examples from recent literature of how decision trees have been applied to analyze epidemiological data.

Summary The overall aim is to encourage epidemiologists to incorporate decision trees into their analytic toolkit.

Keywords Decision trees · CART · Conditional inference tree · Prediction · Subgroups

Introduction

Regression models have been extensively applied in epidemiological research to examine relationships between covariates (e.g., risk factors, demographics, etc.) and an outcome of interest [1, 2]. However, regression models lack the flexibility to uncover complex covariate-outcome relationships unless the analyst pre-specifies the nature of these relationships. For example, consider a randomized controlled trial, the Box Lunch Study (BLS) [3], where one analysis goal was to explore associations between daily food intake measured in kilo-calories and relevant covariates. The set of covariates include responses to the Three-Factor Eating Questionnaire (TFEQ) [4] quantifying hunger, disinhibition, and restrained eating and novel laboratory-based psycho-social measures such as relative reinforcing value of food (rrvf) [5] and degrees of liking and wanting of food [6, 7]. Denoting the outcome by Y and the covariates by $\mathbf{X}$ (for this illustration consisting of six covariates $X_1, \ldots, X_6$), one option would be to fit the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_6 X_6 + \epsilon.$$

This model assumes that daily energy intake is a linear function of covariates, an assumption that is unlikely to hold. One common alternative approach is to categorize continuous covariates, e.g., by splitting them at the median or into quartiles. However, in many cases, it may not be obvious which values to choose for splitting. Furthermore, this approach makes investigating covariate interactions more challenging due to the potentially large number of dummy variables in the model. Investigating different splitting choices requires estimating a potentially large number of linear regression models; when deciding on a single split point for a single continuous X_1, many models would need to be fit to include indicator variables $\mathbf{1}[X_1 \geq \tau]$ defined for different threshold values τ. This may require many candidate models to be examined in a fairly ad hoc manner, potentially inflating type I error and increasing the risk of overfitting the data, which reduces generalizability of the results [8–10].

At the other end of the spectrum, machine learning techniques such as neural networks, support vector machines, and graphical models [11, 12] offer very flexible modeling of covariate-outcome relationships. However, these techniques are usually "black boxes" as they combine covariate

This article is part of the Topical Collection on *Epidemiologic Methods*

✉ Julian Wolfson
julianw@umn.edu

[1] Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Memorial Building MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455, USA

[2] School of Mathematics and Statistics, University of Glasgow, University Place, Glasgow G12 8QQ, UK

information in complex ways. For example, neural networks classify outcomes based on weighted combinations of transformed covariates. The resulting model cannot easily be interpreted in terms of the original covariate values, making it difficult to gain insight about the nature of covariate-outcome associations and make individual predictions without access to software that can calculate model outputs.

Decision trees are an appealing intermediate between these two extremes: they offer more flexibility than standard regression models [13••], but their output is more easily interpretable than "black-box" machine learning methods. As a result, decision trees are potentially useful for analyzing complex, high-dimensional data from epidemiological studies. In this paper, we introduce the key concepts of decision trees and describe how they have been applied in the recent epidemiological literature, distinguishing between three different ways that decision trees are used: for explanatory modeling and variable selection, outcome prediction, and subgroup identification. We also briefly discuss some variants of and extensions to decision tree models which are also seeing increased use in the epidemiological community.

A Brief Overview of Decision Trees

A *decision tree* is a statistical model which aims to partition the given data into groups that are (relatively) homogeneous with respect to the outcome, based on covariate values. Decision trees have several components, which we illustrate by referring to Fig. 1 (note that Fig. 1 is a conditional inference tree, which we describe in greater detail below). Subsets of observations are contained in *nodes* of a decision tree; all observations ($n = 226$ in the Box Lunch Study) are initially contained in the *root node* of a tree (at the top of Fig. 1). The *splitting* step is a vital step in the process of constructing decision trees, where two disjoint subsets are determined by dividing the sample (or subsample, for nodes below the root node) according to covariate values. *Branches* in the tree represent splits below a node; a decision tree is built by successively splitting down each branch until a *stopping rule* is triggered. Stopping rules may be determined by a number of factors, e.g., a minimum number of observations in a node, or a threshold for the decrease in estimated prediction error. A *leaf* or a *terminal node* is a node where the stopping rule is satisfied. Collectively, a disjoint partition of the original sample is defined by terminal nodes; each observation in the sample belongs to a single terminal node, depending on its covariates. Predictions for a new observation are obtained by computing a summary statistic from the individuals falling in the same leaf as that observation. For instance, with continuous outcomes, the predicted value would be the mean outcome of the subset of observations within that leaf. Decision trees are usually depicted upside down relative to actual trees, with the root node at the top and the branches spread downwards to the leaves. The tree in Fig. 1 has four terminal nodes (or "leaves"), and therefore partitions individuals into four subgroups with distinct means (and indeed, distributions) of the outcome on the basis of six variables: hunger, wanting, liking, disinhibition, restrained eating, and relative reinforcing value of food.

The splits in a decision tree define a set of *prediction rules* for predicting the outcome on the basis of covariates, with the goal of minimizing a *loss function* that computes the discrepancy between the predicted and true values. Commonly used loss functions include misspecification rates, Gini index, and entropy for classification trees and mean squared error for regression trees. A *training set* is used to learn a set of decision rules and a *test set* is utilized to assess the performance of the grown decision tree. Like many strongly data-driven methods, decision trees are prone to *overfitting*, i.e., getting an overly optimistic estimate of prediction accuracy by modeling idiosyncrasies of the training set used to build the tree instead of characteristics of the underlying data generating process. To prevent overfitting, it is therefore common practice to construct trees using a number of different stopping rules to generate trees of varying depths (a deeper tree has more splits, more nodes, and fewer observations within each leaf; hence, it may yield higher prediction accuracy but is also more likely to overfit). The final tree depth is selected using a process called *pruning* that seeks to minimize the prediction error estimated by cross-validation or, preferably, on an independent test set.

There are several methods for constructing decision trees, with the major differences between the methods being the algorithms used to partition the sample and the criteria which determine when to stop splitting. The most widely used method of constructing decision trees is the Classification and Regression Tree (CART) technique [15••]. In CART, the search for each split takes place simultaneously across all covariates and their candidate split points. For each covariate, CART identifies the split point resulting in greatest reduction in error. The split chosen for inclusion in the tree is the most error-reducing split across all covariates. This recursive splitting process continues until the best split results in a relative reduction in error less than a pre-specified threshold, since the CART and related techniques (e.g., C4.5 [16]) have seen wide application, including in obesity [17, 18], smoking studies [19•, 20], and diabetes [21, 22].

One more recently proposed alternative to CART is the conditional inference tree (CTree [23]). CTree follows a two-stage splitting process: in the first stage, the covariate to split on is determined based on a measure of association between each covariate and the outcome of interest. Then, the best split point for the splitting covariate is calculated. This two-stage splitting process allows CTree to use a more formal statistical inference framework, wherein the hypothesis that none of the covariates has a univariate association with the

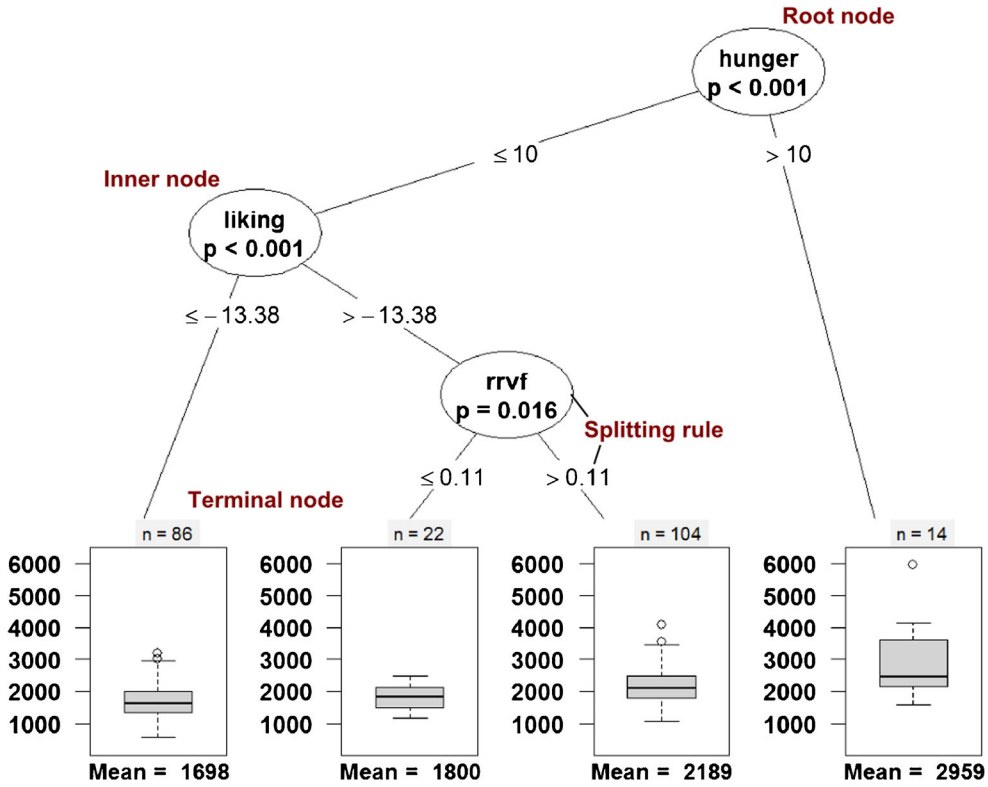

Fig. 1 Conditional inference tree that displays the association between food intake (in kcal/day) and six covariates (hunger, disinhibition, relative reinforcement of food (rrvf), restrained eating, liking, and wanting). The tree represents a series of sequential splits on hunger, liking, and rrvf that distinguish between four subgroups with different distributions of daily caloric intake. Root and inner nodes are labeled with the splitting variable and (multiplicity-adjusted) *p* value for the association between that variable and the outcome. Branches below root and inner nodes are labeled with the optimal splitting rule determined by the CTree algorithm. Terminal nodes display the number of individuals belonging to each subgroup and boxplots showing the distribution of daily caloric intake within each subgroup. This figure is based on a similar one that previously appeared in reference [14] (http://creativecommons.org/licenses/by/4.0/)

outcome is tested by considering a set of tests corresponding to each univariate association and the result of each test is summarized via a *p* value. Nodes are declared as terminal nodes (node IDs 3, 5, 6, and 7 in Fig. 1) when the minimum *p* value determined is larger than a multiplicity adjusted significance threshold. Hence, in conditional inference trees such as the one in Fig. 1, each splitting node is associated with a (multiplicity-adjusted) *p* value and the type I error is controlled both overall and within each node. In a previous paper [14], we compared the advantages and disadvantages of CART and CTree via simulation and found that CART often yielded trees with slightly lower prediction error than CTree but required more parameter tuning, tended to favor the inclusion of continuous over discrete covariates (due to the large number of possible splits of the former), and did not control the overall type I error rate. We argued that the simplicity and inferential focus of conditional inference tress make them an appealing option for epidemiologists, but at this point in time, they have seen limited use in public health and medical research [24, 25•].

Decision trees are widely implemented in open source R statistical software [26] using packages such as rpart [27•] and rpart.plot [28] for CART, partykit [29••] for CTree, and RWeka [30] for C4.5 [16]. Decision trees have also been implemented using SAS Enterprise Miner [31•], in SAS/STAT software using a procedure called hpsplit, and in the CART and CHAID modules in Stata.

Explanatory Modeling and Variable Selection with Decision Trees

One of the advantages of decision trees relative to "black box" machine learning techniques is that they provide interpretable prediction rules in terms of covariates. Hence, they can be used to identify covariates that are most relevant for predicting the outcome. In fact, trees can play two roles in explanatory modeling. First, they can act as a "variable selector" by identifying which available covariates contribute to predicting the outcome. Often, trees are constrained to have a modest number of splits, and hence if the number of available covariates is large, some fraction of those covariates will never appear in the tree and it can be concluded that they do not meaningfully contribute to explaining variation in the outcome. Several

papers we found [32, 33] applied a univariate pre-screening step to identify relevant predictors to include in the tree, but in most situations, we would argue that this is unnecessary since trees already perform the variable selection function described above.

Second, decision trees also play an important role in explaining how covariates influence the outcome. In standard generalized linear regression models, covariates are assumed to be linearly related to some function of the mean. However, relationships may be non-linear so that the effect of a covariate is particularly pronounced over a subset of its range. By constructing data-driven prediction rules based on covariate thresholds, decision trees are better able to detect and characterize such non-linearities. This ability is particularly useful for ordered scales, which are common in clinical contexts and are challenging to handle as either continuous or categorical variables in a regression framework. Though not explicitly designed for effect estimation, a fitted decision tree can be used to estimate the effects of (dichotomized) covariates. For instance, in Fig. 1, the effect of having hunger ≤ 10 vs. > 10 could be estimated by calculating the mean caloric intake twice for every individual: once setting hunger ≤ 10 and once setting hunger > 10, leaving other covariate values fixed. The difference between the two mean caloric intake values is the effect of hunger. Recent papers that have used regression trees for explanatory modeling include Esteban et al. [34], who used CART to identify covariates associated with short-term mortality after an exacerbation of chronic obstructive pulmonary disease (eCOPD). They found that the highest mortality rate was in those with the highest baseline dyspnea level (among 5 levels), and a Glasgow score < 15 (score range 3–15). Kanellos-Becker et al. [35] explored the factors associated with prognosis of midgut volvulus in young children and concluded from a CART analysis that the most important predictors were blood gas analysis base excess (BGA BE) < -1.7 and birth prior to 36 weeks. These factors were then used to derive a prognostic score which had a PPV of 84.2% and an NPV of 100%.

Prediction with Decision Trees

Since trees are a flexible modeling tool, they have been widely applied by researchers seeking to predict outcomes of interest [36•, 37•, 38, 39]. The task of prediction differs from explanatory modeling and variable selection in that, for prediction, the goal is to minimize prediction error, and understanding how covariates contribute to predicting outcomes is less important. How to measure prediction accuracy depends on the outcome type, the available data, and the clinical context. Common metrics include mean-squared error (MSE) for continuous data and area under the ROC curve (AUC) for binary outcome data [40–42]. The most accurate assessment of predictive performance comes when the data is split into separate "training" and "test" sets, with the former used to build the model and the latter used only to assess its performance on independent data [12]. When limited sample size precludes splitting into separate sets, cross-validation is recommended for calculating prediction error. There are several types of cross validation methods including leave-one out cross validation, the holdout method, and *k*-fold cross validation, and software packages in R (e.g., caret [43]) contain built-in cross validation routines that simplify the evaluation of predictive performance.

One drawback of decision trees in the context of prediction is that they can be highly sensitive to small changes in the data. This is because the initial splits of the tree have a major influence on its final structure, and decisions on how and when to split are made on what may be very small differences between the fitting metrics of interest. For example, if splitting a sample on sex reduces the within-subgroup mean squared error (MSE) by 4.7% and splitting it according to whether age is <25 or ≥25 reduces the MSE by 4.6%, then the tree will split on sex. However, it is easy to imagine that a small change in the data might cause the age split to reduce the MSE by 4.8%, in which case age would be chosen to split on instead. This sensitivity is undesirable for prediction, since it means that prediction models based on a tree fitted from one dataset may generalize poorly to new data. To overcome this "sample sensitivity" problem, it is more common to use random forests [44, 45] to derive prediction models. As its name suggests, a random forest is a collection of decision trees, with each tree in the collection fitted from a bootstrap sample of the original data. Final predictions are obtained by averaging the predictions from the trees in the random forest. While random forests often yield more accurate and generalizable prediction models, they lose the interpretability of individual decision trees. Some metrics have been developed that measure "variable importance" within random forests [46–48], but only provide high-level summaries of which variables have the biggest impact; they do not provide insight about specific decision rules and thresholds.

While decision trees are correctly characterized as being less sensitive to underlying assumptions about the relationship between covariates and outcomes than standard regression models, they can predict quite poorly if the outcome scales continuously with covariate values (e.g., the mean of the outcome is a linear function of a continuous covariate). In that case, any covariate cutpoint will produce groups with different means, cutpoints will be essentially arbitrary, and given sufficient data, the tree will split many times. As we have previously shown [14], when the true relationship between covariates and outcomes is linear, decision trees will have much larger prediction error than the standard regression model. Therefore, when contemplating the use of decision trees or random forests to build a predictive model, we strongly recommend comparing their performance to that of an

appropriate generalized linear model to assess whether they offer a meaningful improvement in prediction accuracy.

Subgroup Identification with Decision Trees

Since decision trees are constructed by sequentially splitting the original sample based on covariate values, they classify individuals into distinct population subgroups that are relatively homogeneous with respect to a given outcome. Splits are defined by a set of covariate dichotomizations, so it is straightforward to understand how subgroup membership is determined in a decision tree; this stands in contrast with many classification techniques where the rules used to create subgroups are based on complex rules involving combinations and transformations of covariates and therefore lack a simple scientific interpretation. Decision trees have been used to identify prognostic groups [49], stratify patients in clinical trials [24, 50•, 51], and retrospectively explore treatment/exposure effect heterogeneity [52].

The usual visualization of decision trees (see Fig. 1) includes much of the information needed to characterize population subgroups, but it does not necessarily help researchers comprehend this information. This drawback is pronounced for predictor variables that lack an easily interpretable scale, or when their population distribution is unknown. To address this limitation, as part of our own work, we developed an alternative visualization of the composition of subgroups defined by decision trees [14]. R code for our novel visualization and relevant examples are available at https://github.com/AshwiniKV/visTree. The visualization is presented as a grid of plots, one corresponding to each terminal node (i.e., population subgroup), and summarizes, at a glance, the characteristics of the subgroups identified by the decision tree in Fig. 1. In Fig. 2, a plot is displayed for the terminal nodes (population subgroups) identified by the decision tree in Fig. 1. A histogram shaded in gray is placed in the background of each plot which summarizes the distribution of the outcome variable (here, 24-h energy intake) for individuals that belong to the relevant terminal node/subgroup. The top left plot in Fig. 2 displays a right-skewed distribution of 24-h energy intake and the average 24-h energy intake within each individual bin of the histogram are labeled as numbers above the *x*-axis. The mean and subgroup size for each terminal node are displayed as the plot title and a vertical line shows the overall mean of outcome values contained in the subgroup. Colored bars are overlaid on the background to define the composition of the subgroup; individual bars are placed on the percentile scale to describe the set of predictor values.

The subgroup corresponding to the top left plot of Fig. 2 is defined by liking values below -13.38, which represents the 39th population percentile, and hunger values that are below 10, which represents the 91st percentile. The bottom right plot, by contrast, has left-skewed values of 24-h energy intake and is defined by hunger above 10, where this cut-off point would create the 92nd percentile.

In Fig. 2, the four subgroups are defined by differences in liking, hunger, and relative reinforcing value of food. The first subgroup ($n = 86$) has a below average energy intake (1698 kcal) and is characterized by moderate to low liking and all but very high hunger. The second and third subgroups

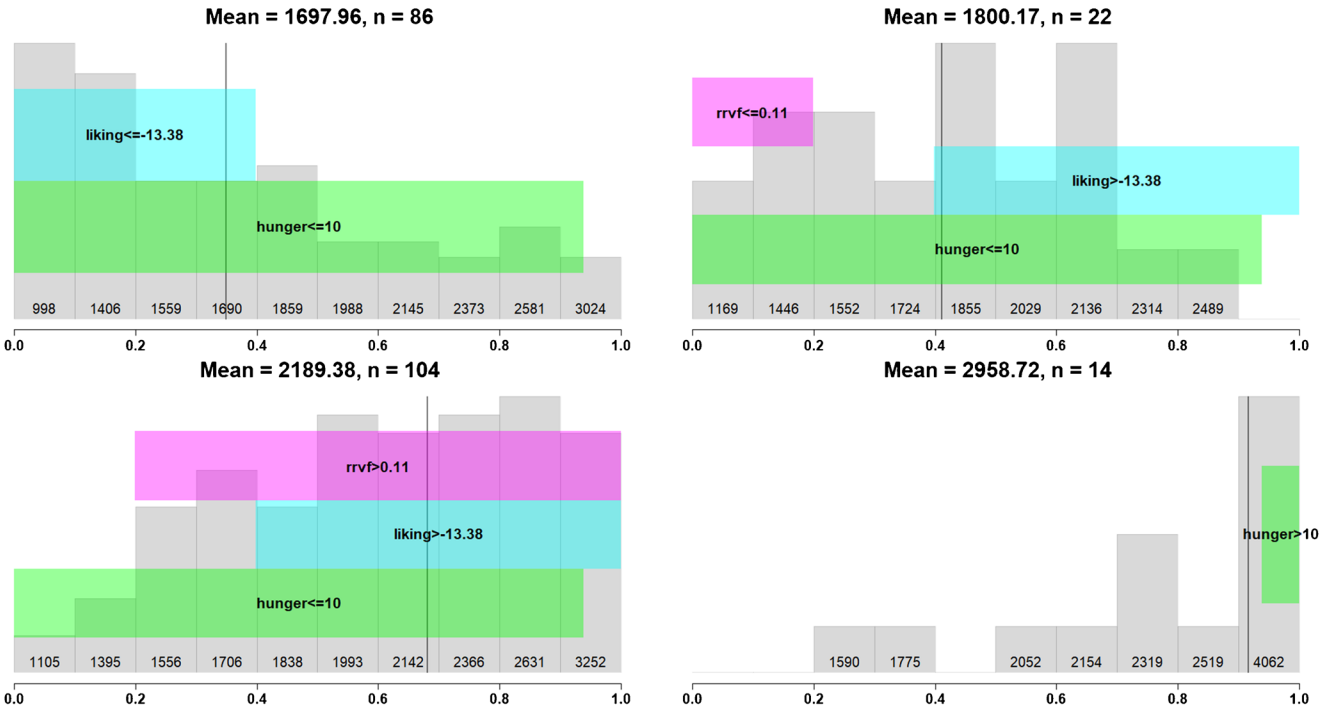

Fig. 2 Visualization that summarizes the characteristics of subgroups identified by the decision tree displayed in Fig. 1

are both characterized by moderate to high liking and all but very high hunger and are differentiated by relative reinforcing value of food; very low rrvf for the second subgroup and all but very low rrvf for the third subgroup. The second subgroup ($n = 22$) has moderate to low energy intake (1800 kcal) and the third subgroup ($n = 104$) has moderate to high energy intake (2189 kcal). The fourth subgroup ($n = 14$) has very high energy intake (2959 kcal) and is characterized by very high hunger.

Conclusion

This article only scratches the surface of an extensive literature on decision trees. The basic concept of sequential dichotomization has been extended in many directions: decision tree variants have been developed to handle virtually all outcome types common to epidemiological studies, including continuous [53], ordinal [54], binary [55•, 56], and time-to-event [57–60] outcomes. Decision trees have been adapted to allow for traditional covariate adjustment and to handle missing covariate data [61, 62•]. They have also been made more flexible by allowing multiway splits (e.g., [63, 64]). As noted above, random forests created by combining predictions from multiple decision trees have yielded very successful and accurate predictors in a wide variety of contexts. Decision trees are also commonly incorporated into other "ensemble" methods which aggregate predictions from a variety of machine learning techniques [65, 66].

Though this article identifies three distinct uses for decision trees, most epidemiologists will choose to apply decision trees because they seek to strike a balance between model complexity and interpretability, and hence will have more than one of these uses in mind. For example, decision trees are an appealing choice when a researcher seeks to build an accurate prediction model that is based on prediction rules that can be implemented in clinical practice. They are also particularly useful for identifying a small number of covariates that can be used to stratify the population into homogeneous subgroups.

With software for fitting decision trees now available in most standard statistical packages, and ongoing work producing visualizations which make the interpretation of decision tree outputs more intuitive, decision trees are starting to be used widely in the scientific literature. We therefore encourage epidemiological researchers to branch out and give decision trees a try.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. Bender R. Introduction to the Use of Regression Models in Epidemiology. In Humana Press; 2009 [cited 2018 Mar 1]. p. 179–95. Available from: http://link.springer.com/10.1007/978-1-59745-416-2_9.
2. Eyler AA, Brownson RC, Bacak SJ, Housemann RA. The Epidemiology of Walking for Physical Activity in the United States. United States Med Sci Sport Exerc [Internet]. 2003 [cited 2018 Mar 1];35(9):1529–36. Available from: https://website.education.wisc.edu/kines119summer/Mod_2_Stuff/EpiWalkingUSA.pdf.
3. French SA, Mitchell NR, Wolfson J, Finlayson G, Blundell JE, Jeffery RW. Questionnaire and laboratory measures of eating behavior. Associations with energy intake and BMI in a community sample of working adults. Appetite [Internet]. 2014 Jan 1 [cited 2018 Mar 1];72:50–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24096082.
4. Stunkard AJ, Messick S. The three-factor eating questionnaire to measure dietary restraint, disinhibition and hunger. J Psychosom Res [Internet]. 1985 Jan 1 [cited 2018 Feb 28];29(1):71–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3981480
5. Epstein LH, Carr KA, Lin H, Fletcher KD, Roemmich JN. Usual Energy Intake Mediates the Relationship Between Food Reinforcement and BMI. Obesity [Internet]. 2012 Sep 13 [cited 2018 Mar 13];20(9):1815–9. Available from: http://doi.wiley.com/10.1038/oby.2012.2.
6. Finlayson G, King N, Blundell J. The role of implicit wanting in relation to explicit liking and wanting for food: implications for appetite control. Appetite [Internet]. 2008 Jan 1 [cited 2018 Mar 13];50(1):120–7. Available from: https://www.sciencedirect.com/science/article/pii/S0195666307003145.
7. Finlayson G, Dalton M. Hedonics of food consumption: are food "liking" and "wanting" viable targets for appetite control in the obese? Curr Obes Rep [Internet]. 2012 Mar 10 [cited 2018 Mar 13];1(1):42–9. Available from: http://link.springer.com/10.1007/s13679-011-0007-2.
8. Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat Sci [Internet]. 2001 Aug [cited 2018 Feb 28];16(3):199–231. Available from: http://projecteuclid.org/euclid.ss/1009213726.
9. Babyak MA. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. Psychosom Med [Internet]. 2004 [cited 2018 Feb 28];411–21. Available from: https://journals.lww.com/psychosomaticmedicine/Abstract/2004/05000/What_You_See_May_Not_Be_What_You_Get__A_Brief,.21.aspx.
10. Roecker EB. Prediction Error and Its Estimation for Subset-Selected Models. Technometrics [Internet]. 1991 Nov [cited 2018 Mar 14];33(4):459–68. Available from: http://www.tandfonline.com/doi/abs/10.1080/00401706.1991.10484873.

11. Cortes C, Vapnik V. Support-vector networks. Mach Learn [Internet]. 1995 Sep [cited 2018 Feb 28];20(3):273–97. Available from: http://link.springer.com/10.1007/BF00994018.
12. Ripley BD. Pattern recognition and neural networks [Internet]. Cambridge University Press; 1996 [cited 2018 Mar 4]. 403 p. s.
13.•• Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. Ann Behav Med [Internet]. 2003 Dec [cited 2018 Mar 24];26(3):172–81. Available from: https://academic.oup.com/abm/article/26/3/172-181/4631556. **This paper provides a review of CART and encourages its utilization towards subgroup identification in public health research.**
14. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. Emerg themes Epidemiol [internet]. 2017 Dec 20 [cited 2018 Feb 28];14(1):11.
15.•• Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and regression trees [Internet]. Chapman & Hall; 1993 [cited 2018 Feb 28]. 358 p. **This paper introduces a popular nonparametric decision tree called the classification and regression tree (CART).**
16. Salzberg SL. C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers, Inc., 1993. Mach Learn [Internet]. 1994 Sep [cited 2018 Feb 28];16(3):235–40. Available from: http://link.springer.com/10.1007/BF00993309.
17. Riedel C, von Kries R, Buyken AE, Diethelm K, Keil T, Grabenhenrich L, Müller MJ, Plachta-Danielzik S Overweight in adolescence can be predicted at age 6 years: a CART analysis in German cohorts. Manco M, editor. PLoS One [Internet]. 2014 Mar 27 [cited 2018 Mar 14];9(3):e93581. Available from: http://dx.plos.org/10.1371/journal.pone.0093581.
18. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Mach Learn Techniques for Prediction of Early Childhood Obesity Appl Clin Inform [Internet]. 2015 Dec 19 [cited 2018 Mar 14];6(3):506–20. Available from: http://www.schattauer.de/index.php?id=1214&doi=10.4338/ACI-2015-03-RA-0036.
19.• Lei Y, Nollen N, Ahluwahlia JS, Yu Q, Mayo MS. An application in identifying high-risk populations in alternative tobacco product use utilizing logistic regression and CART: a heuristic comparison. BMC Public Health [Internet]. 2015 Dec 9 [cited 2018 Mar 14];15(1):341. Available from: http://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-015-1582-z. **This paper discusses the advantages of CART in comparison to logistic regression using an application that seeks to identify target subpopulations.**
20. Nollen NL, Ahluwalia JS, Lei Y, Yu Q, Scheuermann TS, Mayo MS. Adult Cigarette Smokers at Highest Risk for Concurrent Alternative Tobacco Product Use Among a Racially/Ethnically and Socioeconomically Diverse Sample. Nicotine Tob Res [Internet]. 2016 Apr 1 [cited 2018 Mar 14];18(4):386–94. Available from: https://academic.oup.com/ntr/article-lookup/doi/10.1093/ntr/ntv110.
21. Much D, Jaschinski H, Lack N, Hummel S, Füchtenbusch M, Hummel M, et al. Risk Stratification in Women with Gestational Diabetes According to and Beyond Current WHO Criteria. Horm Metab Res [Internet]. 2015 Nov 13 [cited 2018 Mar 14];48(1):16–9. Available from: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0035-1565084.
22. Marinov M, Mosa ASM, Yoo I, Boren SA. Data-Mining Technologies for Diabetes: A Systematic Review. J Diabetes Sci Technol [Internet]. 2011 Nov 1 [cited 2018 Mar 14];5(6):1549–56. Available from: http://journals.sagepub.com/doi/10.1177/193229681100500631.
23. Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework J Comput Graph Stat [Internet]. 2006 Sep [cited 2018 Feb 28];15(3):651–74. Available from: http://www.tandfonline.com/doi/abs/10.1198/106186006X133933.
24. Tanadini LG, Steeves JD, Hothorn T, Abel R, Maier D, Schubert M, et al. Identifying Homogeneous Subgroups in Neurological Disorders. Neurorehabil Neural Repair [Internet]. 2014 Jul 28 [cited 2018 Mar 2];28(6):507–15. Available from: http://journals.sagepub.com/doi/10.1177/1545968313520413.
25.• Cheng FW, Gao X, Bao L, Mitchell DC, Wood C, Sliwinski MJ, et al. Obesity as a risk factor for developing functional limitation among older adults: A conditional inference tree analysis. Obesity [Internet]. 2017 Jul 1 [cited 2018 Mar 14];25(7):1263–9. Available from: http://doi.wiley.com/10.1002/oby.21861. **This paper uses a conditional inference tree to investigate potential risk factors and stratify individuals based on the significant factors.**
26. Team R. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. 2014 [cited 2018 Mar 14]; Available from: https://scholar.google.com/scholar?cluster=8103611549594844363&hl=en&oi=scholarr.
27.• Atkinson, Elizabeth J and Therneau TM. An Introduction to Recursive Partitioning Using the RPART Routines. Rochester Mayo Found [Internet]. 2000 [cited 2018 Mar 14]; Available from: http://r.789695.n4.nabble.com/attachment/3209029/0/zed.pdf. **This report describes methods found in rpart programs to build classification and regression trees (CARTs).**
28. Milborrow S. Rpart.plot: plot 'rpart' models: an enhanced version of 'plot.rpart'. R package version 3.0.0. 2018. Available from: https://CRAN.R-project.org/package=rpart.plot.
29.•• Hothorn T, Zeileis A. Partykit: A Modular Toolkit for Recursive Partytioning in R. J Mach Learn Res [Internet]. 2015 [cited 2018 Mar 14];16:3905–9. Available from: http://www.jmlr.org/papers/volume16/hothorn15a/hothorn15a.pdf. **The partykit package provides a flexible toolkit for prediction, printing and plotting decision trees and a generic infrastructure for recursive partitioning in R.**
30. Hornik K, Buchta C, Zeileis A. Open-source machine learning: R meets Weka. Comput Stat [Internet]. 2009 May 14 [cited 2018 Mar 14];24(2):225–32. Available from: http://link.springer.com/10.1007/s00180-008-0119-7.
31.• Gordon L. Using classification and regression trees (CART) in SAS® Enterprise miner TM for applications in Public Health 2013 [cited 2018 Mar 26]; Available from: http://support.sas.com/resources/papers/proceedings13/089-2013.pdf. **This paper demonstrates applications of CART in public health research using the SAS Enterprise Miner.**
32. Newby D, Freitas AA, Ghafourian T. Pre-processing Feature Selection for Improved C&RT Models for Oral Absorption. J Chem Inf Model [Internet]. 2013 Oct 28 [cited 2018 Mar 12];53(10):2730–42. Available from: http://pubs.acs.org/doi/10.1021/ci400378j.
33. Samanta B, Bird GL, Kuijpers M, Zimmerman RA, Jarvik GP, Wernovsky G, et al. Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms. Artif Intell Med [Internet]. 2009 Jul 1 [cited 2018 Mar 12];46(3):201–15. Available from: https://www.sciencedirect.com/science/article/pii/S0933365708001851.
34. Esteban C, Arostegui I, Garcia-Gutierrez S, Gonzalez N, Lafuente I, Bare M, et al. A decision tree to assess short-term mortality after an emergency department visit for an exacerbation of COPD: a cohort study. Respir Res [Internet]. 2015 Dec 22 [cited 2018 Mar 14];16(1):151. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26695935.
35. Kanellos-Becker I, Bergholz R, Reinshagen K, Boettcher M. Early prediction of complex midgut volvulus in neonates and infants. Pediatr Surg Int [Internet]. 2014 Jun 23 [cited 2018 Feb

28];30(6):579–86. Available from: http://link.springer.com/10.1007/s00383-014-3504-z.

36.• Hostettler IC, Muroi C, Richter JK, Schmid J, Neidert MC, Seule M, et al. Decision tree analysis in subarachnoid hemorrhage: prediction of outcome parameters during the course of aneurysmal subarachnoid hemorrhage using decision tree analysis. J Neurosurg [Internet] 2018 Jan 19 [cited 2018 Mar 13];1–12. Available from: http://thejns.org/doi/10.3171/2017.7.JNS17677. **This paper utilizes decision trees for building prediction models and examining interactions between dependent variables and their influence on the outcome of interest.**

37.• Zimmerman RK, Balasubramani GK, Nowalk MP, Eng H, Urbanski L, Jackson ML, et al. Classification and regression tree (CART) analysis to predict influenza in primary care patients. BMC Infect Dis [Internet]. 2016 Dec 22 [cited 2018 Mar 13];16(1):503. Available from: http://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-016-1839-x. **This paper uses CART to predict an outcome of interest and examine higher order interactions between relevant covariates.**

38. Rendon RA, Mason RJ, Kirkland S, Lawen JG, Abdolell M. A classification tree for the prediction of benign versus malignant disease in patients with small renal masses. Can J Urol [Internet]. 2014 Aug [cited 2018 Mar 13];21(4):7379–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25171283.

39. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control [Internet]. 2017 Feb 9 [cited 2018 Mar 26]; Available from: https://www.sciencedirect.com/science/article/pii/S1746809417300204.

40. Pourahmad S, Hafizi-Rastani I, Khalili H, Paydar S. Identifying Important Attributes for Prognostic Prediction in Traumatic Brain Inj Patients. Methods Inf Med [Internet]. 2016 Aug 5 [cited 2018 Mar 13];55(5):440–9. Available from: http://www.schattauer.de/index.php?id=1214&doi=10.3414/ME15-01-0080.

41. Kojima G, Iliffe S, Tanabe M. Vitamin D supplementation as a potential cause of U-shaped associations between vitamin D levels and negative health outcomes: a decision tree analysis for risk of frailty. BMC Geriatr [Internet]. 2017 Dec 16 [cited 2018 Mar 13];17(1):236. Available from: http://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-017-0631-0.

42. Shi H, Jia J, Li D, Wei L, Shang W, Zheng Z. Blood oxygen level dependent magnetic resonance imaging for detecting pathological patterns in lupus nephritis patients: a preliminary study using a decision tree model. BMC Nephrol [Internet]. 2018 Dec 9 [cited 2018 Mar 13];19(1):33. Available from: https://bmcnephrol.biomedcentral.com/articles/10.1186/s12882-017-0787-z.

43. Kuhn M, Wing J, Weston S, Williams A. The caret Package: Classification and Regression Training R Packag [Internet]. 2015 [cited 2018 Mar 14]; Available from: http://www.download.nextag.com/cran/web/packages/caret/caret.pdf.

44. Tin Kam Ho. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell [Internet]. 1998 [cited 2018 Mar 2];20(8):832–44. Available from: http://ieeexplore.ieee.org/document/709601/.

45. Tin Kam Ho. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition [Internet]. IEEE Comput. Soc. Press; [cited 2018 Mar 2]. p. 278–82. Available from: http://ieeexplore.ieee.org/document/598994/.

46. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput [Internet]. 2017 May 23 [cited 2018 Mar 4];27(3):659–78. Available from: http://link.springer.com/10.1007/s11222-016-9646-1.

47. Hapfelmeier A, Hothorn T, Ulm K, Strobl C. A new variable importance measure for random forests with missing data. Stat Comput [Internet]. 2014 Jan 28 [cited 2018 Mar 4];24(1):21–34. Available from: http://link.springer.com/10.1007/s11222-012-9349-1.

48. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics [Internet]. 2008 [cited 2018 Mar 4];9(1):307. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-307.

49. Lo BWY, Fukuda H, Angle M, Teitelbaum J, Macdonald RL, Farrokhyar F, Thabane L, Levine MAH Aneurysmal subarachnoid hemorrhage prognostic decision-making algorithm using classification and regression tree analysis. Surg Neurol Int [Internet]. 2016 [cited 2018 Mar 14];7:73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27512607.

50.• Tanadini LG, Hothorn T, Jones LAT, Lammertse DP, Abel R, Maier D, et al. Toward Inclusive Trial Protocols in Heterogeneous Neurological Disorders. Neurorehabil Neural Repair [Internet]. 2015 Oct 2 [cited 2018 Mar 14];29(9):867–77. Available from: http://journals.sagepub.com/doi/10.1177/1545968315570322. **This paper describes an application of the conditional inference tree (CTree) where decision rules suggest the participants to be included in a clinical trial and stratify subjects into homogeneous subgroups.**

51. Zhang C, Garrard L, Keighley J, Carlson S, Gajewski B. Subgroup identification of early preterm birth (ePTB): informing a future prospective enrichment clinical trial design. BMC Pregnancy Childbirth [Internet]. 2017 Dec 10 [cited 2018 Mar 2];17(1):18. Available from: http://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-016-1189-0.

52. Tsai W-M, Zhang H, Buta E, O'Malley S, Gueorguieva R. A modified classification tree method for personalized medicine decisions. Stat Interface [Internet]. 2016 [cited 2018 Mar 17];9(2):239–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26770292.

53. Facchinello DY, Beauséjour DM, Richard-Denis DA, Thompson MC, Mac-Thiong DJ-M. The use of regression tree analysis for predicting the functional outcome following traumatic Spinal Cord injury. https://home.liebertpub.com/neu [Internet]. 2017 Oct 25 [cited 2018 Mar 13]; Available from: https://www.liebertpub.com/doi/abs/10.1089/neu.2017.5321.

54. Galimberti G, Soffritti G, Di Maso M. Classification Trees for Ordinal Responses in R : The rpartScore Package. J Stat Softw [Internet]. 2012 [cited 2018 Mar 2];47(10). Available from: http://www.jstatsoft.org/v47/i10/.

55.• Henrard S, Speybroeck N, Hermans C. Classification and regression tree analysis vs. multivariable linear and logistic regression methods as statistical tools for studying Haemophilia Haemophilia [Internet]. 2015 Nov 1 [cited 2018 Mar 4];21(6):715–22. Available from: http://doi.wiley.com/10.1111/hae.12778. **This paper recommends the use of CART in the health domain in comparison to traditional regression models.**

56. Gueorguieva R, Wu R, O'Connor PG, Weisner C, Fucito LM, Hoffmann S, et al. Predictors of Abstinence from Heavy Drinking During Treatment in COMBINE and External Validation in PREDICT. Alcohol Clin Exp Res [Internet]. 2014 Oct [cited 2018 Apr 4];38(10):2647–56. Available from: http://doi.wiley.com/10.1111/acer.12541.

57. Miller RG, Anderson SJ, Costacou T, Sekikawa A, Orchard TJ. Risk stratification for 25-year cardiovascular disease incidence in type 1 diabetes: Tree-structured survival analysis of the Pittsburgh Epidemiology of Diabetes Complications study. Diabetes Vasc Dis Res [Internet]. 2016 Jul 21 [cited 2018 Mar 4];13(4):250–9. Available from: http://journals.sagepub.com/doi/10.1177/1479164116629353

58. Bou-Hamad I, Larocque D, Ben-Ameur H. Discrete-time survival trees and forests with time-varying covariates. Stat Model An Int J [Internet]. 2011 Oct 13 [cited 2018 Mar 13];11(5):429–46.

Available from: http://journals.sagepub.com/doi/10.1177/1471082X1001100503.

59. Last M, Tosas O, Gallo Cassarino T, Kozlakidis Z, Edgeworth J. Evolving classification of intensive care patients from event data. Artif Intell Med [Internet]. 2016 May 1 [cited 2018 Mar 13];69:22–32. Available from: https://www.sciencedirect.com/science/article/pii/S093336571530083X.
60. Linden A, Yarnold PR. Modeling time-to-event (survival) data using classification tree analysis. J Eval Clin Pract [Internet]. 2017 Dec 1 [cited 2018 Mar 13];23(6):1299–308. Available from: http://doi.wiley.com/10.1111/jep.12779.
61. Fortes I, Mora-López L, Morales R, Triguero F. Inductive learning models with missing values. Math Comput Model [Internet]. 2006 Nov 1 [cited 2018 Mar 13];44(9–10):790–806. Available from: https://www.sciencedirect.com/science/article/pii/S0895717706000768.
62.• Bertolet M, Brooks MM, Bittner V. Tree-based identification of subgroups for time-varying covariate survival data. Stat Methods Med Res [Internet]. 2016 Feb 14 [cited 2018 Mar 13];25(1):488–501. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23070595. **This paper presents a variation of the classification and regression tree for survival analysis that also adjusts for potential confounders and accounts for time varying covariates.**
63. Berzal F, Cubero J-C, As Mar In N, Anchez DS. Building multi-way decision trees with numerical attributes. [cited 2018 Mar 13]; Available from: https://pdfs.semanticscholar.org/c907/f7e471ec3b9527ecd0fc1692a4d0f8116e9a.pdf.
64. Kim H, Loh W-Y. Classification Trees With Unbiased Multiway Splits. J Am Stat Assoc [Internet]. 2001 Jun [cited 2018 Mar 13];96(454):589–604. Available from: http://www.tandfonline.com/doi/abs/10.1198/016214501753168271.
65. Parvin H, MirnabiBaboli M, Alinejad-Rokny H. Proposing a classifier ensemble framework based on classifier selection and decision tree. Eng Appl Artif Intell [Internet]. 2015 Jan 1 [cited 2018 Mar 14];37:34–42. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0952197614002048.
66. Chen Y-C, Chen JJ. Ensemble survival trees for identifying subpopulations in personalized medicine Biometrical J [Internet]. 2016 Sep 1 [cited 2018 Mar 14];58(5):1151–63. Available from: http://doi.wiley.com/10.1002/bimj.201500075.