# Data Mining in Psychological Treatment Research: A Primer on Classification and Regression Trees

Matthew W. King VA Boston Healthcare System, Boston, Massachusetts Patricia A. Resick
VA Boston Healthcare System, Boston, Massachusetts, and
Boston University School of Medicine

Data mining of treatment study results can reveal unforeseen but critical insights, such as who receives the most benefit from treatment and under what circumstances. The usefulness and legitimacy of exploratory data analysis have received relatively little recognition, however, and analytic methods well suited to the task are not widely known in psychology. With roots in computer science and statistics, statistical learning approaches offer a credible option: These methods take a more inductive approach to building a model than is done in traditional regression, allowing the data greater role in suggesting the correct relationships between variables rather than imposing them a priori. Classification and regression trees are presented as a powerful, flexible exemplar of statistical learning methods. Trees allow researchers to efficiently identify useful predictors of an outcome and discover interactions between predictors without the need to anticipate and specify these in advance, making them ideal for revealing patterns that inform hypotheses about treatment effects. Trees can also provide a predictive model for forecasting outcomes as an aid to clinical decision making. This primer describes how tree models are constructed, how the results are interpreted and evaluated, and how trees overcome some of the complexities of traditional regression. Examples are drawn from randomized clinical trial data and highlight some interpretations of particular interest to treatment researchers. The limitations of tree models are discussed, and suggestions for further reading and choices in software are offered.

Keywords: classification and regression trees, CART, exploratory data analysis, data mining

There is a persistent demand for researchers to evaluate and improve the techniques used to treat mental illness, driven in part by the progress that still must be made in understanding how, when, and for whom these treatments have beneficial effect. If posttraumatic stress disorder (PTSD) is taken as an example, a body of randomized clinical trial (RCT) data shows that although extant treatments are highly effective on average, the majority of treated patients continue to have substantial residual symptoms and some do not meaningfully respond at all (Bradley, Greene,

This article was published Online First March 3, 2014.

Matthew W. King, Women's Health Sciences Division, National Center for PTSD, VA Boston Healthcare System, Boston, Massachusetts; Patricia A. Resick, Women's Health Sciences Division, National Center for PTSD, VA Boston Healthcare System, Boston, Massachusetts, and Department of Psychiatry, Boston University School of Medicine.

Patricia A. Resick is now at the Department of Psychiatry, Duke University School of Medicine.

This writing was supported by the National Center for PTSD, Office of Patient Care Services, Mental Health Services, Department of Veterans Affairs. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government. The funding agency was not involved in the management, analysis, and interpretation of the data or the preparation, review, or approval of the manuscript.

Correspondence concerning this article should be addressed to Matthew W. King, Women's Health Sciences Division, National Center for PTSD, VA Boston Healthcare System (116B-3), 150 South Huntington Avenue, Boston, MA 02130. E-mail: matthew.king2@va.gov

Russ, Dutra, & Westen, 2005). Taking one approach to advancing the state of the science, Kraemer, Wilson, Fairburn, and Agras (2002) called for researchers to explore RCT data for preliminary evidence of treatment moderators (i.e., baseline individual difference variables that interact with treatment condition to predict outcome: the who and the when) and mediators (i.e., process variables varying across treatment conditions that may help explain observed treatment effects: the how and the why). Their premise was that mining the data of today will generate more rigorous, powerful hypotheses for the clinical research of tomorrow. At the same time, Kraemer et al. lamented the perceived lack of credibility for such data exploration, attributing this to a widespread insistence that a confirmatory framework is the only proper mode of analysis. The fact that advanced exploratory techniques are scarcely covered in most graduate psychology statistics curricula (Aiken, West, & Millsap, 2008) may also be partially to blame.

Many psychology researchers, as a cause or a consequence, likely associate the notion of exploratory data analysis (EDA) only with the rather dry (and often rushed) precursors to substantive analyses—computing sample descriptives, checking regression assumptions, and the like. But in contrast, the founder of the EDA tradition, mathematician John Tukey (1977), championed the approach as productive in its own right. Like a detective, the researcher engaged in EDA observes the facts as they are, constructs meaningful new hypotheses from discovered patterns of data, and uses his or her existing knowledge to identify those plausible enough to submit to future confirmatory work (Behrens & Yu, 2003; Tukey, 1977). True EDA creates the fuel for confirmatory

research, not by generating answers but by generating stronger questions that lead to more carefully designed and powerful studies, more return on research dollar investment, and more rapid and efficient progress in treatment approaches.

# Complexities of Traditional Regression for Exploratory Analysis

Even those researchers who do on occasion mine their datasets for novel and interesting patterns (perhaps only behind closed doors) may perceive few analytic options besides the traditional tools of confirmatory regression: ordinary least squares (OLS) linear or logistic regression, repeated measures analysis of variance (ANOVA), and the like. When researchers use EDA however, exclusive reliance on such techniques entails some complexities that may not be fully appreciated.

#### **Restrictive Assumptions**

Traditional regression makes assumptions about the shape of the population underlying the data and about the form of the relationships between variables (cf. Cohen, Cohen, West, & Aiken, 2003). When OLS regression is used, for instance, results depend on the assumption of a straight-line relationship (i.e., each unit change in a predictor is accompanied by a constant change in outcome) and may be seriously misleading if the true relationship is otherwise. In exploratory analysis, where learning about the form of predictor–outcome relationships is often a specific interest, a priori assumptions such as linearity may be undesirable. It is often possible to transform variables in order to model nonlinear forms, but doing so makes interpretation more difficult and introduces additional analytic complexities.

#### **Number of Predictors**

During EDA, the researcher will likely wish to compare the importance of several predictors simultaneously. In traditional regression, the number of predictors that can enter a model simultaneously is strictly limited by the sample size in order to ensure stable estimates (Cohen et al., 2003). For many study designs and perhaps especially in treatment research, sample sizes tend to be small due to logistical issues (e.g., difficult recruitment, expensive study protocol), but each subject provides copious data, leading to a situation where the researcher has many more candidate predictors to test than can possibly be accommodated simultaneously. This limitation can be partially mitigated by using statistical criteria to determine which predictors enter the model (i.e., stepwise regression), but in common algorithms, the entry criteria most typically test for an additive linear effect of new predictors and do not (automatically) consider the possible contributions of interactions.

# **Interaction Terms**

The sample size limit becomes even more constraining when the researcher decides to test for interactions between predictors. Each interaction term counts as an additional predictor against the overall total the sample can support, so it is rarely feasible to examine more than a few two-way interactions at once, let alone include higher order interactions. An additional complication is

that the interactions to be examined (or any that will compete for entry) must be specified prior to fitting the regression model. This requirement can undermine an exploratory analysis, where the goal is the discovery of unexpected interactions.

### "Statistical Learning" as a Complement to Regression

Fortunately, recent advancements in statistical computing have expanded the options accessible to researchers interested in data mining. With roots in computer science and statistics, statistical learning refers broadly to approaches that take a more inductive approach to building a model, allowing the data greater role in suggesting the correct relationships between variables rather than imposing them a priori (Hastie, Tibshirani, & Friedman, 2009). A number of different statistical learning methods can produce regression-like findings while avoiding some of the complexities of traditional approaches, making them ideally (although not exclusively) suited to exploratory analysis. The current paper focuses on one family of statistical learning models known as classification and regression trees. Following sections will describe how a tree model is constructed and how it is evaluated and its results interpreted. In the spirit of full disclosure, some of the limitations of tree models will be discussed as well. Analysis examples will be presented to better illustrate the principles underlying tree models. These will be based on data from an exemplar clinical study (Resick, Nishith, Weaver, Astin, & Feuer, 2002): an RCT where 171 female rape survivors with PTSD were treated with cognitive processing therapy (CPT; Resick & Schnicke, 1993) or prolonged exposure (PE; Foa & Rothbaum, 1998).

Statistical learning is an active, diverse area of research, and new refinements of tree methods emerge in the literature frequently. This presentation will be general when possible and will highlight the principles common to all trees, but it will also focus specifically at times on the CART algorithm (Breiman, Friedman, Olshen, & Stone, 1984). The reasons are primarily pedagogical; CART is widely considered a progenitor of today's modern tree methods and is the baseline against which more refined methods are often compared. Many additional introductory and intermediate level expositions on CART are widely available for the reader interested in learning more about tree models, and, ultimately, a solid understanding of CART facilitates learning the particulars of its methodological descendants.

#### **Classification and Regression Trees**

To show how tree methods build a model of the data and why the model is useful for predicting outcomes and describing relationships between variables, we start with a cartoon example of basic regression principles. Suppose you worked at a carnival and it was your job to guess the weights of carnivalgoers. In order to perform well for "the house," you must find a way to minimize the total amount by which you are off in your guesses over the run of all the guesses you make. If you were required to make your

<sup>&</sup>lt;sup>1</sup> Although it is an acronym for *classification and regression trees*, CART actually refers to one specific algorithm for building trees, rather than pertaining to these types of models more broadly, and is a registered trademark of California Statistical Software, Inc. We will use "CART" only when referring to the specific algorithm.

guesses without having any information about the persons—you didn't see or hear them or even know their names—you would be obliged simply to guess the mean weight of the population of carnivalgoers every single time, because this value minimizes your accounting of errors (commonly a function like the sum of squared prediction errors; i.e.,  $\sum [weight_{true} - weight_{guess}]^2$ ). You can outperform this naive guess over the long run, however, if you have some additional information about the persons that is correlated with their weights, like their heights. Knowing that people who are similar in height are more likely than people on the whole to be similar in weight, you will have a smaller overall accounting of errors if, for each person, you guess the mean weight only of people of similar height. This is called the *conditional mean* of weight given height. If you also know the person's gender, you can shrink your error further by guessing the mean weight only of men or women of similar height, as appropriate (i.e., mean weight conditioned on height and gender). Examining how conditional means change across different predictor values is the essence of regression analysis (Berk, 2008), for ordinary linear regression and tree methods alike. The goal in using conditional means to minimize prediction errors is the same in both, but whereas linear regression assumes the conditional means all fall on a straight line (i.e., a unit increment in height always predicts the same amount of conditional weight increase, at all points along the height scale), tree methods allow the pattern to be dictated by the data.

# A Tree-Building Algorithm

Computer algorithms are used to construct a tree model of a regression problem. An algorithm is simply a problem-solving procedure that involves a series of systematic steps. At each step, one or more rules are evaluated to determine how (or whether) to proceed further (think of a flowchart that involves decision points and different branches that a process might follow). What follows is the essence of the steps that tree-building algorithms follow. In practice, each of these steps (save the first) is carried out by a computer program.

Step one: Assign variables to outcome and predictor roles. From the available variables, the researcher selects a single outcome (response, dependent) variable and a number of predictors. The outcome can be ordered (ordinal, interval, ratio) or nominal (dichotomous or polytomous). Predictors may also be either ordered or nominal. Tree methods can theoretically consider as many predictors as desired; the limiting factor is primarily computer processing time.

To illustrate, we will mimic a categorical prediction problem originally described by Rizvi, Vogt, and Resick (2009). Looking at patients in the Resick et al. (2002) RCT for PTSD, Rizvi et al. attempted to predict treatment dropout (a dichotomous outcome, dropout, coded as true if the patient did not complete at least 75% of psychotherapy sessions) using six baseline individual difference variables: the patient's age in years (age), years of education (education), estimated intelligence quotient (IQ), and levels of pretreatment depression (depression), trait anger (anger), and global feelings of guilt (guilt). Rizvi et al. analyzed these predictors separately and each in interaction with the randomized treat-

ment condition (tx\_cond; CPT or PE), but with a tree, all seven can be considered simultaneously.

Step two: Evaluate all of the predictors to find the "best" binary split of the outcome variable. A binary split creates two subgroups at a given cutpoint. For ordered predictors, the two subgroups comprise subjects falling on either side of a cut score; for example, for IQ, possible cut scores include 90, 100, or 110, where patients with an IQ below the cut constitute one group and all other patients constitute the second group. For nominal predictors, cuts are between all different binary groupings of the classes; for example, if tx\_cond comprised the conditions CPT, PE, and wait list (WL), cutpoints would include CPT versus PE/WL, PE versus CPT/WL, and WL versus CPT/PE.

The object of splitting the sample is to create two subgroups that are more homogenous or internally similar with respect to the outcome than the sample on the whole. For example, if patients with below-average intelligence are more likely to drop out of treatment, then splitting on IQ between average and below-average ranges will create a group below the cut with a higher proportion of dropouts than the group above the cut.

Different tree methods vary in their approach to identifying the "best" binary split. The archetypical CART algorithm performs an exhaustive evaluation of all cutpoints across all predictors, measuring the extent to which each cutpoint is able to reduce outcome heterogeneity, also referred to as impurity. When the outcome is quantitative, it is common to calculate impurity within a group with a familiar measure of variance, the mean squared error (MSE; i.e.,  $\sum [X - M]^2/N$ ). When the outcome is categorical, an impurity function based on relative class proportions is often used. A typical choice is the Gini index, which in the current example with two classes (dropouts and nondropouts) can be defined as p(1-p), where p is the proportion of dropouts in a group. On a plot, the Gini index resembles an inverted *U*: It reaches a minimum when p is 0 or 1 (no impurity) and a maximum when p = .5 (maximally heterogeneous). The single split that offers the greatest impurity reduction between the larger sample and the subgroups wins the competition.

Step three: Split the sample and repeat step two within each subgroup. With the best split selected, divide the sample into two subgroups according to the given cutpoint. Within each subgroup, repeat the process starting at step two. That is, treat each grouping as a new sample in its own right and restart the competition between predictors. Every predictor remains up for consideration again, including the predictor last selected. Continue evaluating predictors and splitting into progressively smaller groups until some set criteria for stopping are met.

What if there are missing data among the predictors? Among the several compensatory approaches, the strategy used by CART is illustrative (if not optimal). When a predictor is initially evaluated, CART uses only observed values to determine its optimum split. But if the predictor is selected, CART will not have the information it needs to assign subjects with missing values to the appropriate subgroup. CART attempts to overcome this problem by using *surrogate predictors*. In a case of CART-within-CART or "CART-lite" (Berk, 2008), subjects with observed values on the winning predictor are split into two subgroups accordingly, and this classification temporarily becomes a new outcome variable. All potential splits on other predictors are evaluated, and the predictor and cutpoint able to sort subjects into this outcome

grouping with the fewest errors are chosen. For subjects with missing values on the original winner, values on the winning surrogate are used instead to assign them to a subgroup (unless surrogate values too are missing, in which case second-place surrogate values are used, and so on). CART does not permit missing values on the outcome variable.

To see the tree-building process in application, consider the scatterplot shown in Figure 1, which depicts the association among IQ, guilt, and dropout. The association of IQ with dropout across the whole sample is negative: The proportion of dropouts (compared to the proportion of completers) generally decreases with increasing IQ. A split on IQ at a cutpoint of 91.5 (vertical single dashed line) successfully captures a relatively high proportion of dropouts to the left of the cut ( $\leq$ 91.5) and a relatively low proportion of dropouts to the right ( $\geq$ 91.5). Splitting can continue within these two subgroups, and, within the lower IQ group, there is another negative association between guilt and dropout: Again, the proportion of dropouts decreases with increasing guilt. A split on guilt at a cutpoint of 2.125 (horizontal double dashed line) separates a relatively high

proportion of dropouts below the cut (<2.125) and a relatively low proportion of dropouts above the cut ( $\ge 2.125$ ). Note that the same association is not present in the higher IQ group and that no horizontal line would be especially effective at sectioning off a group of similar outcomes. This suggests that guilt is unlikely to be selected for the winning split among patients with an IQ above 91.5, but some other predictor may usefully subdivide this group further.

Figure 1 helps to show why CART and related methods are referred to as *recursive partitioning* algorithms: The effect of applying a series of splits to the sample is to create progressively smaller, increasingly homogenous, rectangular partitions of the data, and this is achieved with a process that repeats itself on inputs defined by previous iterations (a kind of programming logic called *recursion*). Different tree methods vary in the "stopping rule" utilized to decide when splitting is halted, and these can also usually be manipulated by the researcher according to the analysis goals. With CART, it is typical for the process to be halted when splitting would reduce subgroup size below a predetermined minimum limit.

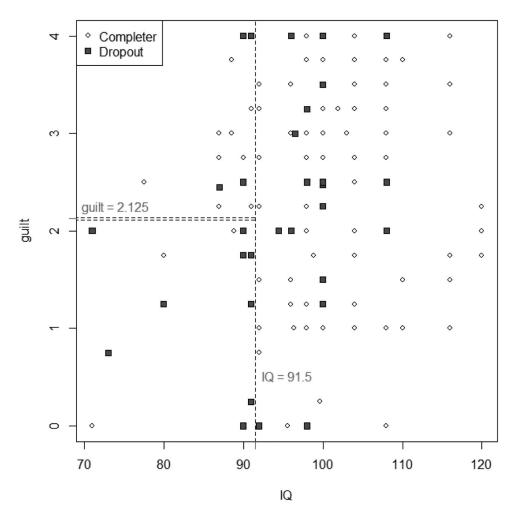


Figure 1. Scatterplot of IQ, pretreatment global guilt scores, and treatment dropout status. The first partition (single dashed vertical line) represents an initial optimal split on IQ, and the second partition (double dashed horizontal line) represents a subsequent optimal split on guilt among patients with lower IQ.

## **Interpreting a Tree Model**

Figure 2 depicts CART output and demonstrates why the model is called a tree: The typical visualization resembles an inverted tree, with a single root at the top leading to a number of branches and finally several nonbranching leaves at the bottom. Along with the root, each "stop" along the tree is called a *node*. Internal nodes correspond to splits in the sample, with the node label denoting the predictor that was used and the two branches indicating where in the scale of the predictor the split occurred. In the nonbranching leaves or *terminal* nodes, the splitting process was terminated, and the bar plots denote the relative proportions of each outcome (dropout vs. completer). The size of the terminal nodes is also shown.

The tree in Figure 2 depicts a CART model of the Rizvi et al. (2009) dropout prediction problem. At the root node, all patients who completed at least one psychotherapy session (n = 145) are evaluated, and the winning split was in IQ. Patients with an IQ greater than or equal to 91.5 (n = 115) follow the left branch, whereas patients with an IQ less than 91.5 (n = 30) follow the right. Among the latter group, the next winning split was in guilt. Sixteen patients with guilt scores greater than or equal to 2.125 were separated from 14 patients with lower levels of guilt. The resulting partitions are fairly homogenous: 71% of lower guilt patients dropped out as compared to only 25% of those with higher guilt. Patients with higher IQ were also split further: First, 95 patients who were at least 22.5 years old were split from 20 younger patients, and, finally, in the younger group, 10 patients with anger scores below 20 were split from 10 patients with higher levels of anger.

**Outcome variable predictions.** As in traditional regression, tree models provide *fitted* values for sample members: model

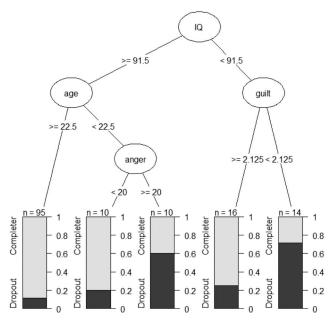


Figure 2. Classification tree produced by CART analysis of the predictors of treatment dropout. Nodes with branches denote binary splits of the sample using the indicated predictor, with branches indicating the cutpoint. Bar plots in nonbranching terminal nodes indicate the proportion of dropouts with particular combinations of relevant predictor values.

predictions of a subject's likeliest outcome given the subject's scores on the predictors. These values are generated by fitting some sort of prediction model in each of the terminal nodes. Again, different tree methods vary in the terminal models used, but in CART these are elegantly simple: CART attempts to minimize its total accounting of prediction errors, and in many classification tree applications (such as Figure 2) this simply means assigning subjects to the majority class in their respective terminal node. As an illustration, imagine "dropping" a given patient down the tree from the root node and following the appropriate branches until a terminal is reached. For example, a patient with an IQ of 100 who is 18 years old and has a pretreatment anger score of 30 would end in the third terminal node, where 60% of similar patients dropped out of treatment. Therefore this patient would be classified by the tree as a dropout. Similarly, a patient with an IQ of 90 and a pretreatment guilt score of 3 would end in a node where only 25% of similar patients dropped out. Therefore this patient would be predicted to complete treatment.

This exercise should begin to reveal why tree models are popular in the medical literature: When each branching node represents some clinical sign or symptom and the object is to predict some prognostic outcome, the tree model becomes a useful aid to clinical decision making (cf. Harper, 2005). Psychological treatment researchers frequently face prognostic questions as well. As with the guess-the-weight example, if you were asked to predict which patients in this sample were likely to drop out of treatment but had no information about them, you would be obliged to predict zero dropouts; this is the only sensible strategy for minimizing your overall prediction errors given that dropouts are a minority, even though your error rate with true dropouts would be 100%. On the other hand, predicting dropout for patients with certain combinations of predictor values (i.e., lower IQ and lower guilt or higher IQ but younger age and higher anger) cuts this error rate with dropouts almost in half and achieves a lower overall error rate as well. Note that this was achieved without any necessary foreknowledge of which predictors would be useful for predicting dropout (beyond selecting the predictors to be considered). A later section discusses how this improvement in prediction performance was calculated and explores other methods for assessing the goodness-of-fit of a tree model.

Interactions and main effects. One of the great strengths of tree models is their ability to reveal unforeseen interactions between predictors. If one considers the model in Figure 2, the appearance of guilt in the right branch of IQ but not the left represents an interaction: the effect of guilt on dropout depends on the level of IQ. Age also interacts with IQ, and anger is part of a triple interaction with IQ and age. Recall that these interactions did not have to be specified in advance, as would be the case with a traditional regression model. Thus, use of tree models allows the researcher to discover localized conditional effects (e.g., the importance of anger among patients with higher IQ but lower age) that may otherwise have been missed.

It should be noted, however, that the advantage is a double-edged sword: A tree model is unlikely to depict a statistical main effect, which is the most common effect studied with traditional regression. To understand the reason for this, consider that a main effect in a typical regression model is essentially a statistical hypothesis; the researcher imposes the main effect on the data and, when absent evidence of significant

interaction with other predictors, opts for parsimony and treats the main effect as a valid representation of reality. But in real datasets, it is rarely the case that the effect of a predictor on the outcome is truly identical at all levels of another substantive predictor. Even if true in the population, fluctuations due to sampling error bias observed data away from such a finding. Because trees reflect empirical rather than imposed relations between variables, these fluctuations influence which predictor and cutpoint is selected for a given split, making it unlikely that exactly the same predictor and cutpoint will be selected in both branches from a node or that the magnitude of two conditional effects will be identical (see Strobl, Malley, & Tutz, 2009, for further discussion).

Another example will help illustrate and also make other important points. Suppose we were interested in learning whether changes in the use of particular coping strategies over the course of therapy were associated with the treatment response. Patients in the Resick et al. (2002) RCT completed the Coping Strategies Inventory (CSI; Tobin, Holroyd, & Reynolds, 1984) at baseline and posttreatment, which measures eight specific strategies they might use when coping with the aftermath of sexual assault. How best to represent change in two-wave data is a debated topic, but for simplicity, raw change scores ( $X_{\rm post} - X_{\rm pre}$ ) in coping strategy use were used to predict raw change in PTSD symptomatology (Clinician Administered PTSD Scale scores; Blake, Weathers, Nagy, & Kaloupek, 1995). The observed RCT data were artifi-

cially manipulated slightly for this example to make the key points more clearly.

Consider the CART tree in the left panel of Figure 3, which utilized all eight strategies plus tx\_cond as predictors. With a continuous outcome, terminal nodes comprise distributions of PTSD change scores; in CART, fitted values are determined by a constant prediction model, typically the conditional mean within each terminal. Results suggest that the amounts by which patients reduced their use of social withdrawal (soc\_withdraw) and selfcriticism (self\_crit) are associated with treatment response. The model shows an interaction between these predictors, but follow-up visual analysis can help determine if the additional model complexity (i.e., an interaction instead of two main effects) appears needed. In the left panel of Figure 4, change in self\_crit is plotted against change in PTSD, with different symbols used for patients above and below the -7.5 cutpoint on soc\_withdraw. Locally weighted smoother (lowess) lines suggest the form of the association. The effect of soc\_withdraw is evident in the distance between lowess lines. But it also seems that the association between self\_crit and PTSD is quite different between groups; a tentative interpretation might be that reductions in self-criticism have limited impact if social withdrawal does not also decline.

The absence of tx\_cond outwardly suggests the model fits within both groups, but here the interpretation is not straightforward because posttreatment predictors may be correlated with treatment assignment. When predictors are correlated, they may

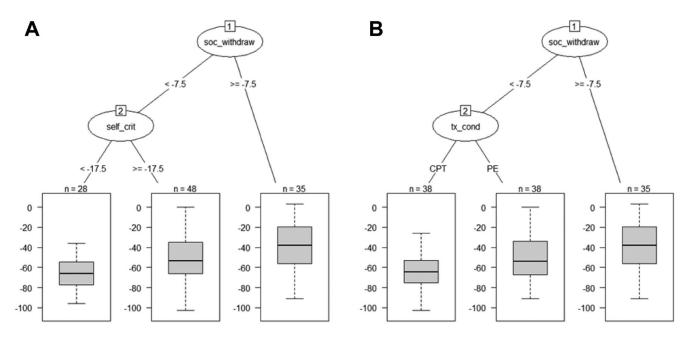


Figure 3. Regression trees produced by CART analysis of change in PTSD symptom severity during a psychotherapy randomized clinical trial, using as predictors the change (growth or decline) in coping strategy use during the treatment period. Boxplots in the terminal nodes indicate the distribution of symptom change scores for particular combinations of predictor values (the bold horizontal lines denote the median scores, the boxes denote the interquartile range, and the whiskers denote the minimum and maximum scores). The tree in the left panel (A) includes all predictors, whereas the tree in the right panel (B) excludes the self-criticism predictor (self\_crit). The selection of the randomized treatment condition (tx\_cond) as a predictor when self\_crit is unavailable suggests that self-criticism and treatment group may be correlated and competing for similar splits on PTSD change scores. PTSD = posttraumatic stress disorder; soc\_withdraw = social withdrawal; CPT = cognitive processing therapy; PE = prolonged exposure.

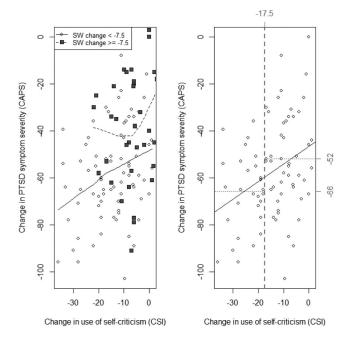


Figure 4. Scatterplots of PTSD symptom severity change against change in the use of self-criticism as a coping strategy. The left panel depicts two groups created by a split in a second predictor: change in use of social withdrawal (SW). The interaction between self-criticism and social withdrawal is revealed by the differences between the smoothed trend lines between groups. The right panel (SW change <-7.5) depicts the larger group, the subsequent split in self-criticism selected by CART, and the two piecewise fitted values used to model the prediction function. PTSD = posttraumatic stress disorder; CAPS = Clinician Administered PTSD Scores; CSI = Coping Strategies Inventory.

compete to split the outcome in closely similar ways. When the analysis is rerun with self\_crit excluded, the model in the right panel of Figure 3 results: With its competitor out of the running, this time tx\_cond is selected for the split. One tentative interpretation is that reductions in self-criticism partially account for the (fictional) between-group difference. Of course mediation should not be inferred from these results alone, as other associations could give rise to the same split decisions. Tree results, although rich and enlightening, can be challenging to interpret and require follow-up to further understand the relationships between predictors. For example, examining the "runners-up" for a particular split can be helpful in identifying when predictors provide a similar level of impurity reduction in the outcome, and a proper path analysis can help determine if a true mediating relationship exists.

Form of the prediction function. The Figure 3 example is useful for highlighting another insight into the way trees model complex relationships. A traditional regression model is "linear in its parameters," meaning the individual effects (however transformed) are combined additively to form a straight-line equation used to calculate a fitted outcome value (cf. Cohen et al., 2003). In a tree model, each terminal node is like a categorical effect (e.g., a dummy variable) in a regression equation, predicting (in CART) a constant fitted outcome value for a given set of predictor values. An implication of this is that the overall prediction function, which describes how the fitted outcome value changes with a change in

predictor values, is constructed one piece at a time. Consider the scatterplot in the right panel of Figure 4, which plots the larger soc\_withdraw group (<-7.5) from the left panel. The linear regression line predicting change in PTSD from self\_crit for these data is plotted (solid line). Overlaid are the two constant fitted values from the CART model (-66 for self\_crit <-17.5, -52 for self\_crit  $\ge -17.5$ ; horizontal dotted lines). Note that the linear regression and the tree are both trying to characterize the same predictor—outcome relationship, and this represents the tree's best approximation with a single binary split (although either subgroup could in theory be split again on self\_crit if the data warranted it).

There are two important points to consider. One is that, as shown in Figure 4, despite the power of trees to discover relationships empirically, the fit of a tree model is not always superior to the fit of a traditional regression model, if in fact the latter is a more appropriate reflection of the population relationships underlying the data. The prudent researcher is advised to continue exploring the data visually to assess the tenability of tree results and to consider more parsimonious models when feasible. Second, on the other hand, a piecewise prediction function has considerable flexibility. It is able to model relationships that change slope or even direction, which are complexities that traditional regression can capture only with considerable prespecification of the necessary effects.

#### Assessing the Fit of a Tree

Just as with traditional regression, the adequacy of fit of a tree model depends on the strength of the associations between the predictors and the outcome. Given the parallel goals of trees and traditional regression (i.e., minimization of prediction errors), methods available to assess the fit of a tree model should appear familiar.

**Fit of a regression tree.** With a regression tree, it is typical for each subject to be assigned a constant predicted value according to the conditional mean within his or her respective terminal node. The difference between each subject's actual outcome value and the predicted value can be treated as a customary residual (i.e., error of prediction,  $Y - \hat{Y}$ ) and analyzed with the same tools common to traditional regression. The average residual (i.e., root mean square error) is

$$RMSE = \sqrt{\frac{\sum (Y - \hat{Y})^2}{N}}$$

Also known in linear regression as the standard error of the estimate, the RMSE quantifies the average accuracy of the tree's predictions; the smaller the RMSE, the more homogenous the terminal nodes in the tree model and the better the fit of the model to the data. It is also possible to use residuals to calculate a measure known as *proportional reduction in error* (PRE). Recall that the sum of squared deviations around the sample mean outcome,  $\sum (Y - M)^2$ , or total sum of squares (TSS), is a measure of total sample variance. The residual sum of squares or RSS,  $\sum (Y - \hat{Y})^2$ , is a measure of variance remaining after predicting outcomes using the tree model. Thus, PRE can be calculated as a ratio of explained to total variance, or (TSS - RSS)/TSS. This provides a proportion between 0 and 1 and can be interpreted as an  $R^2$  or variance explained measure. For both trees depicted in

Figure 3, the RMSE was 20.8 (in the same units as the outcome measure, so points on the CAPS scale in this case) and the PRE was .20.

**Fit of a classification tree.** With a classification tree, subjects are assigned to a predicted class (often the majority class within their respective terminal node), and for each subject, that predicted class is either correct or incorrect. Thus, a researcher may use customary measures for assessing the performance of a classifier to quantify the fit of the tree model. The top half of Table 1 shows the  $2 \times 2$  contingency table, or *confusion matrix*, for the classification tree in Figure 2. Cells on the diagonal hold the numbers of correct predictions (for both classes: completers and dropouts), whereas off-diagonal cells hold the numbers of misclassifications. A prediction error rate for a particular class can be calculated by dividing the number of misclassifications of that class by the number of class members. For example, the tree incorrectly predicted treatment completion for 17 of 33 patients who dropped out, for an error rate of .52. In contrast, the tree misclassified only .07 of treatment completers (8 of 112). This yields an overall error rate of (8 + 17)/145 = .17. Note that if one particular class can be designated as the "positive instance" of interest, familiar concepts like sensitivity and specificity can be used. In this case, the tree has a sensitivity of .48 and a specificity of .93 for classifying dropouts.

In addition to making class predictions, researchers may wish to make use of the relative class proportions in the terminal nodes. For example, referring to Figure 2, it may be meaningful to note that 71% of patients with lower IQ and lower guilt and 60% of patients with higher IQ but lower age and higher anger dropped out of treatment, rather than simply noting that both combinations were classified as dropouts. Class proportions carry more information about the fit of certain predictor combinations and may be useful for further analysis; for instance, class proportions have been used as propensity scores for predicting subjects' selection of treatment condition in group designs that do not involve random assignment (Luellen, Shadish, & Clark, 2005). Class proportions also allow for another quantitative assessment of class discriminability, using the familiar area under a receiver operating characteristic (ROC) curve (AUC; Hanley & McNeil, 1982). The AUC in this case is equivalent to the probability that a randomly selected dropout falls into a terminal node with a higher proportion of dropouts than present in the terminal node of a randomly selected

Table 1 Confusion Matrices for Two CART Analyses of the Predictors of Treatment Dropout

Predictor	Completion predicted	Dropout predicted	Model error rate
	Equal error costs $(FN = FP)$		
Completed treatment	104	8	.07
Dropped out of treatment	17	16	.52
Overall error rate			.17
	Unequal error costs (FN = $2*FP$ )		
Completed treatment	90	22	.20
Dropped out of treatment	10	23	.30
Overall error rate			.22

Note. FN = false negative (misclassifying a true dropout as a completer); FP = false positive (misclassifying a true completer as a dropout).

completer. An AUC approaching 1.0 indicates that most patients fall into a terminal node with a strong majority for their given class (i.e., good separation of classes into homogenous partitions), whereas an AUC close to .5 indicates that most terminal nodes retain considerable heterogeneity (and thus discriminability is poor). In Figure 2, the AUC was .75, suggesting fair discriminability between dropouts and completers.

It is not uncommon to see performance that heavily favors the majority class when the minority class (as in this example) is comparatively rare. We alluded earlier to the fact that CART assigns predicted classes in a way that minimizes its accounting of errors; thinking of each error as having some level of "cost," CART wants to pay the least total cost across all the errors it makes. When (as is typically true by default) all types of errors are treated as having equal cost, minimizing errors with the more common class is the least costly strategy. The researcher whose application involves errors with differing relative costs may wish to change this default behavior, an approach discussed in detail by Berk (2008). For example, we might decide that false negatives (missing those who do drop out) are twice as costly as false positives (misclassifying those who stay in treatment); the former involves a lost opportunity to provide additional support that could keep a patient in treatment, whereas the latter involves extra attention to the patient that is not strictly necessary. Most tree algorithms can be instructed to account for these costs in each split decision (equivalent to weighting dropouts as if they occurred more often in the population than in the sample). Refitting the model with these new costs yields the confusion matrix shown in the bottom half of Table 1. The overall error rate has increased slightly (.22 vs. .17) and the error rate with completers has increased substantially (.20 vs. .07), but the error rate for dropouts is much improved (.30 vs. .52). The structure of the resulting tree will no longer have a straightforward interpretation, but this approach may be useful if the researcher is looking for insights relevant to clinical decision making.

Tree model size and fit. All else being equal, larger trees provide a better fit. Consider again the right panel of Figure 4; the larger the tree, the more pieces are used to construct the prediction function, and the more closely it can adapt to the data. But also, the larger the tree, the smaller the lowermost nodes become, and splitting decisions become increasingly influenced by a small number of (often extreme) cases. Thus, larger trees risk overfitting when the model has started adapting to "noise" (i.e., samplespecific random variation that does not reflect systematic associations in the population). Overfitted trees tend to have unstable, nonreplicable structures and are generally poor forecasters for new data from the same population. But even large trees that are not overfitted tend to be more difficult to interpret and to lose some of their practical value. For example, the tree in Figure 2 suggests the very testable hypothesis that clinicians who address patients' anger and guilt levels prior to starting trauma-focused therapy may be able to increase the completion rate, but a larger, more complex tree may involve so many variables that deriving concise, testable hypotheses for future studies becomes exceedingly difficult.

Thus, all tree methods should strike a size balance between improving fit versus improving generalizability and interpretability, and various methods can be used to this end. Many methods attempt to prevent unstable splits by enforcing a minimum node size. Some methods, such as CART and the more advanced

GUIDE (Loh, 2002), utilize *pruning*, which is much as it sounds: Starting with the leaves of a fully grown tree, pruning removes undesirable splits (i.e., the resulting subgroups are rejoined) to produce a simpler subtree. Pruning can be guided by a "penalty" charged for each terminal node; if a given split does not result in an accuracy improvement in excess of the penalty paid, the split is removed. The penalty term can be varied in order to influence the size of the pruned tree. The logic is the same as when penalized fit criteria (e.g., Akaike information criterion, Bayesian information criterion) are used to limit the number of predictors in stepwise regression.

Pruning is often implemented in conjunction with a fit assessment strategy called cross-validation. As a general rule, any measure of fit will be overly optimistic if calculated from the same data used to grow the tree (called the training dataset) as opposed to an independent sample of the same variables (a test dataset). Crossvalidation attempts to mimic test data with these steps: (a) the sample is randomly divided into m equal-sized, nonoverlapping subsamples (10 is a frequent default); (b) each of the *m* subsamples is held out in turn while a tree is grown from the other m-1subsamples; (c) the predictive accuracy of each tree is tested with the one subsample not included in its training set; and (d) averaging over the m fit assessments. This method can be used to determine a reasonable value for the pruning penalty term; that is, results from cross-validation identify the penalty needed in order to prune splits from the full tree that do not prove reliable across the *m* tests. The remaining splits are less likely to be the products of random variation in a subsample or a few extreme cases.

Some tree methods do not rely on cross-validation and pruning at all but instead control tree size through sophisticated stopping rules. The conditional inference tree (CTree; Hothorn, Hornik, & Zeileis, 2006) algorithm makes split decisions using p values from statistical hypothesis tests (adjusted for multiple comparisons if desired) between the predictors and outcome variable in each node. Splitting terminates when no null hypothesis of independence can be rejected at a prespecified  $\alpha$  level. This rule helps control overfitting by avoiding splits based on chance-level variation in the data. Trees can be made steadily smaller by enforcing stricter criteria (e.g., higher  $\alpha$  for each test).

# **Advancements in Tree Models**

It is instructive to consider CART as an archetype tree method, but it is also important to understand its limitations, as this facilitates an introduction to the tree-building refinements seen in more contemporary approaches.

Unbiased variable selection. CART performs an exhaustive evaluation of possible splits at each node, basing its variable selection on the impurity reduction criterion. A consequence of this is that predictors with a greater number of distinct values (e.g., continuous vs. discrete) are preferentially selected (Loh, 2011; Strobl, Boulesteix, & Augustin, 2007), due to competitive advantage in the number of potential cutpoints and thus opportunities for (possibly trivially) superior impurity reduction. This can obviously mislead researchers on the true substantive importance of the various predictors. To avoid this bias, contemporary methods tend to rely on statistical hypothesis tests for variable selection. For example, GUIDE evaluates the chi-square test of independence for each predictor with the outcome and selects the predictor with the

smallest p value, and then performs a cutpoint search on impurity reduction for that predictor only. CTree uses permutation tests to select the winning predictor. These approaches more reliably identify the predictor with the strongest conditional association with the outcome and are less susceptible to chance variation. As an added refinement, some methods can also select winners based on hypothesis tests of the association between the outcome and interactions of predictors or linear combinations of predictors, to avoid an exclusive focus on univariate main effects that might obscure the conditional importance of some predictors (Loh, 2009).

Smoother prediction function. As noted earlier, CART (among other methods) constructs prediction functions as piecewise constants. This gives CART flexibility but also makes it more difficult to represent the shape of smooth functional forms (like a linear slope). Such forms also present difficulty in modeling appropriate changes in fitted outcome values for cases at the hard boundary near a cutpoint. Consider, for example, the first split in either tree in Figure 3. In practical terms, there is unlikely to be a meaningful difference between a decline in use of social withdrawal coping of 8 points versus 7 points, but patients just on either side of that hard boundary are placed in groups with mean PTSD symptom score declines of 58 points versus 38 points—a clinically significant difference. Subjects near cutpoints will often systematically not be fit well by piecewise-constant functions. Some contemporary methods (such as GUIDE) ameliorate this weakness in fit by using more sophisticated and adaptive terminal node prediction models, such as simple linear or polynomial regressions of terminal node outcome values on the best available predictor (Loh, 2011).

**Random forests.** Despite the many refinements, single trees risk instability under some circumstances; that is, they may exhibit susceptibility to changing structure under small changes of the data. As noted, this risk increases with overfitted trees and small samples, but also when available predictors are only weakly related to the outcome and/or some predictors are too strongly correlated with each other. In all cases, split selection is tenuous and tends not be replicable (say, in a cross-validation sample or new data). But even with adequate data, an intransigent issue is that tree methods involve "greedy" algorithms: At each stage, a locally optimal decision about how to split the sample is made, without any knowledge of future stages. Thus, trees do not always arrive at a globally optimal solution that maximizes prediction accuracy. In particular, there may be predictors highly useful for predicting outcome in a small partition of the sample, but a single tree may never find these if preceding splits do not make this partition available. The researcher may be left with an incomplete sense of this predictor's importance.

Some limitations of a single tree can be overcome by using what are called ensembles or committees of trees, such as a *random forest* (Breiman, 2001). The essence of a random forest is that a single tree algorithm is applied repeatedly to grow a large number of trees (say, 500). Each tree is grown from a bootstrap sample of the data (i.e., random sampling with replacement), and, at each stage, only a randomly selected subsample of predictors compete to provide a cutpoint for a split. This has the effect of making the trees very diverse and gives predictors that would have been overlooked in a single tree much greater opportunity to be selected. Fitted values are determined by combining across trees, commonly by averaging quantitative outcomes or going with majority vote on

categorical outcomes (e.g., a patient is classified as a dropout if more than 50% of trees predict this class). In this way, random instabilities in individual trees tend to cancel out, producing more reliable predictions. Further, the great diversity of trees (and thus of fitted values across subjects) tends to produce much smoother prediction functions (think about how a sequence of straight line segments, if divided small enough, begin to resemble a smooth curve). In typical applications, only the trees where a subject was not included in the training sample contribute to the fitted value for that subject (i.e., the average or the vote, as appropriate). As such, random forests incorporate the idea of cross-validation and produce more honest assessments of fit and forecasting ability. Trees that can be combined as ensembles include CART, CTree, and GUIDE.

Despite the many strengths of a random forest, its Achilles' heel is that interpretation becomes significantly more complicated. There is no longer a single tree that can be examined to learn how the predictors relate to the outcome, requiring the use of indirect methods (e.g., partial dependence plots for visualizing aspects of the prediction function; Hastie et al., 2009). Comparisons between predictors on predictive importance are currently limited to ordinal ranking only (Strobl et al., 2009). Thus, the main utility for random forests lies in producing comprehensive (and, as is usually preferable, more honest) assessments of model fit and more reliable, less biased subject outcome predictions. Random forests can also be useful if a researcher truly has a rather large number of predictors and needs an efficient method to identify a subset worthy of closer examination. But for exploratory data analysis, where gaining insight into the interaction between predictors is usually a high priority, random forests may supplement single trees but probably do not replace them.

#### Conclusion

Our goal with this brief primer has been to provide a conceptual introduction to classification and regression trees and, we hope, to foster interest in this credible and productive approach to exploratory data analysis. The primary strength of trees is the efficient identification of useful outcome predictors and unanticipated relationships between variables. Trees are flexible, making few assumptions of the data, allowing a variety of variable types to be used as outcomes or predictors, and employing a built-in mechanism for working around missing predictor values. And they do this without extensive analytic planning and data preparation needed from the researcher. These characteristics make trees ideal for discovering clues to the whos, whens, and hows of treatment effects that can be tested in future studies.

It is important to view tree models within their proper context, however. Tree modeling complements traditional regression analysis but does not replace it. Traditional regression models will often have better fit when they are properly specified and their assumptions are met, and they remain necessary for testing certain effects. Researchers should avoid any inclination to treat trees like a mechanical truth detector. Trees can be led astray by weaknesses in the data or the algorithm. Results should always be filtered through the lens of subject matter knowledge for plausibility, and they should not be considered definitive without independent confirmatory study.

By necessity, many important topics have been left to further reading, not least of which is the "how to" of using tree modeling software. Researchers will find that a variety of software packages implement some flavor of tree modeling. The popular analysis programs SPSS and SAS offer add-on modules that provide the ability to construct tree models. The author of GUIDE offers free software and documentation for this algorithm for download (http:// www.stat.wisc.edu/~loh/guide.html). Among other free of charge options, arguably the de facto standard implementations for CART and CTree are available in the statistical programming software R. All analyses and graphics presented here are CART trees produced in R (R Core Team, 2012), using packages rpart (Therneau & Atkinson, 1997) and partykit (Hothorn & Zeileis, 2012). CTree is available through the package party (Hothorn et al., 2006). It should be noted that researchers entirely accustomed to point-andclick analysis software may find the learning curve for R rather steep, but some excellent published reference books exist. Everitt and Hothorn's (2010) chapter on conducting tree analyses in R is recommended, although the novice user will probably wish to supplement with an introductory text. Good tutorials on both tree modeling and R in general abound on the Internet.

The interested reader will likely benefit from consulting intermediate level references on this topic before undertaking substantive classification and regression tree analysis. We refer to Berk (2008), Merkle and Shaffer (2011), and Strobl et al. (2009) for excellent choices. Of these, the paper by Strobl et al. is likely to be the most accessible to behavioral scientists, but all delve more comprehensively "under the hood" of these algorithms than was possible here and provide source lists for the growing number of published studies in psychology and social science that utilized tree models. We believe that researchers in psychology, in both treatment research and beyond, will value an investment made in adding this powerful technique to their methodological repertoire.

#### References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, 63, 32–50. doi:10.1037/0003-066X.63.1.32

Behrens, J. T., & Yu, C. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 33–64). Hoboken, NJ: Wiley.

Berk, R. A. (2008). Statistical learning from a regression perspective. New York, NY: Springer.

Blake, D. D., Weathers, F. W., Nagy, L. M., & Kaloupek, D. G. (1995).
The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress*, 8, 75–90. doi:10.1002/jts.2490080106

Bradley, R., Greene, J., Russ, E., Dutra, L., & Westen, D. (2005). A multidimensional meta-analysis of psychotherapy for PTSD. American Journal of Psychiatry, 162, 214–227. doi:10.1176/appi.ajp.162.2.214

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. New York, NY: Chapman & Hall. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Everitt, B. S., & Hothorn, T. (2010). A handbook of statistical analyses using R (2nd ed.). Boca Raton, FL: Chapman & Hall.

- Foa, E. B., & Rothbaum, B. O. (1998). Treating the trauma of rape. New York, NY: Guilford Press.
- Hanley, J. A., & McNeil, M. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36
- Harper, P. R. (2005). A review and comparison of classification algorithms for medical decision making. *Health Policy*, 71, 315–331. doi:10.1016/ j.healthpol.2004.05.002
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning (2nd ed.). New York, NY: Springer.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15, 651–674. doi:10.1198/106186006X133933
- Hothorn, T., & Zeileis, A. (2012). partykit: A toolkit for recursive partytioning (Version 0.1-4) [Software package]. Retrieved from http://CRAN.R-project.org/package=partykit
- Kraemer, H. C., Wilson, G., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. Archives of General Psychiatry, 59, 877–883. doi:10.1001/ archpsyc.59.10.877
- Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12, 361–386.
- Loh, W.-Y. (2009). Improving the precision of classification trees. Annals of Applied Statistics, 3, 1710–1737. doi:10.1214/09-AOAS260
- Loh, W.-Y. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1, 14–23.
- Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review*, 29, 530–558. doi:10.1177/0193841X05275596
- Merkle, E. C., & Shaffer, V. A. (2011). Binary recursive partitioning: Background, methods, and application to psychology. *British Journal of Mathematical and Statistical Psychology*, 64, 161–181. doi:10.1348/000711010X503129

- R Core Team. (2012). R: A language and environment for statistical computing (Version 2.15.1) [Software package]. Vienna, Austria: R Foundation for Statistical Computing.
- Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. *Journal of Consulting and Clinical Psychology*, 70, 867–879. doi:10.1037/0022-006X.70.4.867
- Resick, P. A., & Schnicke, M. K. (1993). Cognitive processing therapy for rape victims: A treatment manual. Newbury Park, CA: Sage.
- Rizvi, S. L., Vogt, D. S., & Resick, P. A. (2009). Cognitive and affective predictors of treatment outcome in cognitive processing therapy and prolonged exposure for posttraumatic stress disorder. *Behaviour Re*search and Therapy, 47, 737–743. doi:10.1016/j.brat.2009.06.003
- Strobl, C., Boulesteix, A.-L., & Augustin, T. (2007). Unbiased split selection for classification trees based on the Gini index. *Computational Statistics & Data Analysis*, 52, 483–501. doi:10.1016/j.csda.2006.12.030
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Meth*ods, 14, 323–348. doi:10.1037/a0016973
- Therneau, T. M., & Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines (Tech. Rep. Series No. 61). Rochester, MN: Mayo Clinic.
- Tobin, D. L., Holroyd, K. A., & Reynolds, R. V. C. (1984). *User's manual for the Coping Strategies Inventory*. Athens: Ohio University, Department of Psychology.
- Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Received December 18, 2012
Revision received November 25, 2013
Accepted December 5, 2013