Published in final edited form as:

Acad Radiol. 2007 March ; 14(3): 306–311. doi:10.1016/j.acra.2006.12.013.

Use of Classification and Regression Trees in Diuresis Renography¹

José Nilo G. Binongo, PhD, Andrew Taylor, MD, Andrew N. Hill, PhD, Brian Schmotzer, MS, Raghuvver Halkar, MD, Russell Folks, CNMT, Eva Dubovsky, MD, PhD, Ernest V. Garcia, PhD, and Amita K. Manatunga, PhD

Department of Biostatistics, Rollins School of Public Health, 1564 Clifton Road N.E., Emory University, Atlanta, GA 30322 (J.N.G.B., A.N.H., B.S., A.K.M.); Department of Radiology, Division of Nuclear Medicine, School of Medicine, Emory University, Atlanta, GA (A.T., R.H., R.F., E.V.G.); and Department of Radiology, School of Medicine, University of Alabama, Birmingham, AL (E.D.).

Abstract

Rationale and Objectives—Decision support systems have the capacity to improve diagnostic performance and reduce physician errors. The purpose of this study was to evaluate the use of classification and regression trees (CART) with bootstrap aggregation as a decision support system in the baseline plus furosemide (F + 20) diuresis renography protocol to determine when obstruction can be excluded without the furosemide acquisition and to identify the key parameters for making this determination.

Materials and Methods—Patients with suspected ureteral obstruction were randomly assigned to a training set (80 patients, 157 kidneys) and a validation set (64 patients, 124 kidneys). Forty quantitative parameters (curve parameters, MAG3 clearance and voiding indices) were generated from each baseline Tc-99m mercaptoacetyl triglycine (MAG3) scan. Three expert readers independently evaluated each kidney regarding the need for furosemide and resolved differences by majority vote. CART with bootstrap aggregation was applied to the training set to generate prediction algorithms which were tested in the validation set.

Results—The algorithm agreed with the expert decision on the necessity of furosemide in 90% (111 of 124 kidneys), with misclassification rates of 10.0% and 10.9% for the left and right kidneys, respectively. The most important discriminators were the postvoid-to-maximum count ratio, the cortical 20-minute-to-maximum count ratio, and the postvoid-to-1-to-2-minute count ratio.

Conclusion—CART can identify the key parameters for discriminating between nonobstruction and possible obstruction, has the potential to serve as a decision support tool to avoid unnecessary furosemide imaging, and can be applied to more complex imaging problems.

Keywords

CART; bootstrap; bagging; misclassification rate; diuresis renography; Tc-99m mercaptoacetyl triglycine; MAG3; radionuclide renography; decision support systems; expert systems

¹This research was supported by grant R01 LMN07595.

© AUR, 2007

Address correspondence to A.K.M.: amanatu@sph.emory.edu.

Decision support systems have the capacity to improve diagnostic performance and reduce physician errors. In particular, decision trees present an attractive way of summarizing expert knowledge for convenient use by nonexperts. Decision trees provide a simple flowchart prescription of a short series of yes/no questions that result in a decision relevant to the scientific question of interest. These ideas have been in use in the social sciences since the 1960s and 1970s (1, 2). The statistical use of these concepts was developed in 1984 by Breiman et al. (3), who named the method “classification and regression trees,” now more commonly known by its acronym CART. Although this procedure has wide applicability, it is known to be sensitive to small changes in the data. Recently, a modified version of CART called “bootstrap aggregation” (or “bagging”) was developed; this new approach has been reported to stabilize the variability of the standard CART procedure (4).

In the evaluation of suspected ureteral obstruction, a recent international consensus panel recommended baseline radionuclide imaging with Tc-99m mercaptoacetyltriglycine (MAG3) followed by furosemide administration and an additional 15 minutes of imaging (5). The authors of the consensus report noted that baseline imaging might be sufficient to exclude obstruction in experienced hands (5), and some investigators have recommended specific strategies to determine when the furosemide acquisition can be omitted (6). This approach avoids unnecessary imaging, minimizes patient inconvenience, and reduces costs by saving technologist, camera, computer, and physician time. Nevertheless, selective elimination of the furosemide acquisition was not the general recommendation of the consensus report because physicians performing low volume studies may not have had the training or experience to make this determination with confidence. The purpose of this study was to evaluate the use of CART as a decision support system to determine when furosemide is required to exclude obstruction and to identify the most important kidney parameters for making this determination.

MATERIALS AND METHODS

Study Populations

Analysis of the training set and validation dataset was performed under the purview and with the approval of the institutional review board. The study was HIPAA compliant, and informed consent was not required. Renal studies from 80 randomly selected patients (33 men and 47 women, mean age 57.5 [SD 18.0] years, 79 right and 78 left kidneys) were used as a training set. The validation set consisted of 64 randomly selected patients (36 men and 28 women, mean age 53.6 [SD 15.1] years, 64 right and 60 left kidneys). All studies used for this development were obtained from the renal database of patients referred for suspected renal obstruction.

Acquisition Protocol and Data Collection

Patients were positioned supine, with the scintillation camera detector placed under the table. A three-phase dynamic baseline acquisition was begun at the time of injection of approximately 10 mCi of Tc-99m MAG3. Phase one consisted of twenty-four 2-second frames, phase two was sixteen 15-second frames, and phase three was forty 30-second frames. The processing of all patient studies was based on the QuantEM 1.0 renal quantification program (7); an updated version of the software, QuantEM 2.0, was used to generate the input parameters. The QuantEM^M software, developed specifically for Tc-99m MAG3, incorporates several quality control procedures to improve reproducibility, generates specific quantitative parameters recommended for scan interpretation, and allows the MAG3 clearance to be calculated using a camera-based technique.

Briefly, for the baseline renogram, a static image is summed from the 2- to 3-minute postinjection frames. Using a filtered version of this image, whole kidney, background, and cortical regions of interest (ROIs) are automatically defined using a thresholding technique. The user can override any of these automatic ROIs and replace them with manual ROIs. The size of the kidney ROI is based on the number of pixels contained within the ROI. Background-subtracted whole kidney and cortical curves are generated and 40 quantitative parameters are generated including curve parameters (time to peak counts and 20-minute-to-maximum count ratios [20 min/max ratio] for both whole kidney and cortical ROIs), voiding indices consisting of the postvoid-to-1-to-2-minute (postvoid/1-2 min ratio), postvoid-to-prevoid, and postvoid-to-maximum (postvoid/max ratio) count ratios; relative percent uptake (relative uptake); and a camera-based MAG3 clearance. The camera-based MAG3 clearance is calculated from the 1- to 2.5-minute whole kidney uptake of MAG3 corrected for renal depth and attenuation and the preinjection and postinjection images of the dose syringe (7); this method of calculating the MAG3 clearance has been previously validated in a multicenter trial (8).

Expert Panel Review

The readers were defined as “expert” based on the fact that each reader had more than 20 years experience in full-time academic nuclear medicine, had multiple publications in renal nuclear medicine, and had been invited to give renal nuclear medicine educational sessions at national radiology and national nuclear medicine meetings. The three expert readers, blinded to the outcome, independently scored each kidney on the need for furosemide and resolved differences by majority vote; the majority decisions were used as the decisions of the expert panel for comparison with the decisions support generated by CART. Based on the decision of the expert panel, 81 of 157 kidneys in the training set required furosemide and 56 of 124 kidneys in the validation set required furosemide. The prediction algorithms were built up from a training set using a 20-parameter subset for each kidney as described above.

Standard CART Procedure

CART produces a single decision tree based on the data from the training set, as illustrated in Figure 1. The methodological details of decision trees are well described by Breiman et al. (3). The basic element of the tree is a node, which can either be an internal node or a terminal node. In Figure 1, internal nodes are represented by circles; terminal nodes are represented by rectangles.

A tree is grown in a hierarchical manner. At each internal node, a binary (yes/no) question is asked. As a first step, at the top node, the algorithm finds the parameter among all kidney parameters and the cutoff point among all possible cutoff points that does the best job of differentiating between kidneys that need and those that do not require furosemide to exclude obstruction. This process of splitting at each node is continued until a large tree is constructed. The large tree usually overfits the data (i.e., being overly sensitive to irregularities in data). An over-fitted tree runs the risk of correctly predicting the outcome for all subjects in the training set, yet ending up so specifically tailored to the training set that it performs poorly on many other datasets. A pruning rule is thus implemented to determine the proper tree size. As a final step, a misclassification rate is calculated in each terminal node of the tree.

CART With Bagging

A potential problem with using a single tree (as in standard CART) on which to build a prediction model is that small perturbations in the training data can result in drastically different trees. Errors made in an early dichotomy are passed down to subsequent splits, thus

compounding the error. In order to stabilize the algorithm, k classification trees are constructed by the common statistical technique of bootstrapping the training data (9), where k is a large number. In brief, bootstrap sampling is a process that randomly selects a single kidney from the training set, assigns that kidney to the bootstrap dataset, randomly selects another kidney from the training set (this kidney could potentially be the same as the first kidney), assigns that kidney to the bootstrap dataset and continues this process until a bootstrap sample the same size as the original training set has been constructed. This whole process is then repeated for k iterations to produce k bootstrapped datasets. A tree (algorithm) is developed for each of the k bootstrapped datasets to determine the need for furosemide. These k algorithms are applied to each kidney in the validation data, resulting in k predictions for each kidney. The final prediction regarding furosemide is determined by simple majority vote of the k outcomes. This methodology, called bootstrap aggregation or bagging (4), reduces dependence on the training set and stabilizes the prediction algorithm by averaging the results. k is chosen to be an odd number (in this study, 1001) of bootstrap samples to avoid any ties in voting. Descriptive statistics are calculated for the k bootstrapped trees in terms of their misclassification rates. This bootstrap procedure also allows an assessment of the variability of the misclassification rate of single trees. The R package (10) was used for implementing the procedures in this paper, and the code is available upon request. Computations were performed on a personal computer. Statistical procedures for CART are available in the S-PLUS environment (11) and in the R freeware system (10, 12).

RESULTS

When the standard CART procedure was applied to the right kidney ($n = 79$) in the training set, Figure 1 is obtained. As the figure shows, the original tree for the right kidney consisted of four levels with six terminal nodes. In this particular tree, four kidney parameters were found to be useful: postvoid/max ratio, relative uptake, cortical 20 min/max ratio, and MAG3 clearance. Whether or not a patient's kidney needs furosemide depends on the kidney's values for these variables. For example, a right kidney that has a postvoid/max ratio < 0.3781 , relative uptake $< 86\%$, and cortical 20 min/max ratio < 0.5084 is predicted not to require furosemide. In the training set, 34 right kidneys had this set of characteristics. The number 0/34 in the leftmost terminal node of Figure indicates the misclassification rate for this particular path, which was 0. For this particular tree, the total misclassification rate was 2.5% (2 of 79).

When this tree was applied to the right kidneys ($n = 64$) in the validation sample, the misclassification rate was 15.62% (10 of 64) (Table 1); 1 of 29 (3.45%) right kidneys needing furosemide were misclassified, and 9 of 35 (25.71%) not requiring furosemide were misclassified. Because the data in the validation set were not used in building the original tree, the increase in the number of misclassified kidneys was expected.

When bagging was applied, the modified CART algorithm reduced the misclassification rate for the right kidney from 15.62% to 10.94% ($p = 0.03$, based on McNemar test), misclassifying 2 of 29 (6.90%) who needed furosemide and 5 of 35 (14.29%) of those who did not need furosemide. The misclassification rates for single trees ranged from 4.69% to 35.94% (Table 1), indicating large variability for single trees. The bagging misclassification rate for the right kidney was also smaller than the mean (or median) misclassification rate of the bootstrapped samples (Table 1). These comparisons suggest that bagging stabilized standard CART analysis.

The same procedure was performed on the 78 left kidneys in the training set. When the original tree was applied to the validation sample consisting of 60 left kidneys, 3 of 27

(11.11%) of patients needing furosemide were misclassified and 2 of 33 (6.1%) of patients who did not need furosemide were misclassified, resulting in an overall misclassification rate of 5 of 60 (8.33%). The bagging misclassification rate of 4 of 27 (14.81%) for patients needing furosemide and 2 of 33 (6.06%) for patients who did not need furosemide with an overall misclassification rate of 6 of 60 (10.00%). The bagging and original tree misclassification rates were not significantly different ($p = 0.32$). Thus, for the left kidney, bagging did not provide an advantage over the original tree as far as reducing the number of misclassified cases. Again, however, bagging stabilized the standard CART algorithm; without bagging, the misclassification rate ranged from 6.7% to 30.0%, whereas the misclassification rate with bagging was 10% (Table 1).

As Figure 1 shows, a tree can have more than one level. Quite a few bootstrapped trees, however, have only one level, and 74% and 88% of the trees for the right and left kidneys, respectively, have a maximum of three levels (Table 2). The three most frequently selected kidney parameters on level 1 for both right and left kidneys were (a) postvoid/max ratio, the ratio of the counts in the kidney after voiding urine to the maximum counts in the kidney, (b) cortical 20 min/max ratio, the ratio of the counts in the cortex of the kidney at 19-20 minutes using a parenchymal (cortical) region of interest to maximum counts in the same cortical region of interest, and (c) postvoid/1-2 min ratio, the ratio of the counts in the kidney after voiding urine to counts at 1 to 2 minutes (Table 3).

DISCUSSION

It is generally accepted that experts interpret studies in their specialty better than general radiologists; this is the basis for having distinct areas of expertise within academic departments and private practice settings. The goal of expert systems is to help physicians interpret studies with the same level of expertise as experts. Expert systems have the capacity to improve diagnostic performance and reduce physician errors, particularly in low-volume studies where physicians may have had limited training and experience. Although classification trees have long been proposed as tools for decision making, they have recently come into their own with the widespread accessibility of significant computing power and appropriate software.

CART with bagging accurately predicted the need for furosemide to discriminate between nonobstruction and possible obstruction about 90% of the time. The algorithm also specified the parameters used at the various levels of the bootstrapped sampled trees to determine when obstruction could be excluded without the furosemide acquisition; this analysis provided an important insight into the parameters that are most important in discriminating between obstruction and non-obstruction. The time to half-peak ($T_{1/2}$), for example, is frequently cited as an important measurement in evaluating possible obstruction (14), but this was not an important variable in the CART analysis for determining the need for furosemide (distinguishing between nonobstruction and possible obstruction). In fact, two of the three most frequently selected parameters at the first level employed a comparison of the counts in the kidney after voiding to an earlier time period (maximum counts or counts at 1-2 minutes) (Table 3) and support an earlier study suggesting that voiding indices will provide simple and more robust parameters for evaluating obstruction than the $T_{1/2}$ (13).

An attractive feature of CART is that the same kidney parameter may occur at more than one level in a particular tree, although the parameter will be associated with a different cutoff point. This feature allows for additional flexibility in the algorithm. CART is a nonparametric method, making no assumptions about underlying distributional behavior. This allows for analysis of many datasets, but parametric modeling may provide viable alternatives to this approach. Although not an issue in this work, CART can handle patients

with missing data without completely discarding the observations for such patients, and it is designed to handle a large number of variables. The flexibility is further illustrated by the fact that results obtained using CART are unaffected by the choice of measurement units and reexpression of variables. The use of CART is attractive for its simplicity and in the speed with which a set of prediction outcomes can be obtained.

Bagging stabilizes the conventional CART procedure. Hastie et al's remark (4, p. 247) that single trees may have high variance due to the high correlation in the predictors is applicable to the present study. Bagging smoothes out this variance and reduces the misclassification rate. Perhaps the most significant disadvantage in the bagging technique is the lack of a simple tree at the end of the procedure on which to base future predictions. The final prediction is too complicated to be presented visually. CART with bagging cannot provide the interpreting physician with a rationale for reaching a specific conclusion. Nonetheless, we maintain this is a price worth paying in light of the reduction in variability and possible subsequent improvement in overall misclassification rate of the bagged tree. Another obvious limitation of the study is that the results only apply to the baseline plus furosemide diuresis protocol and will not apply to protocols where furosemide is administered before or simultaneously with the tracer.

This paper did not account for the within-subject correlation structure. It is known that the variability of prediction can be minimized by making adjustments for possible correlation between each patient's left and right kidneys (15). However, in this study the odds ratio relating the need for furosemide in the right and left kidneys was weak ($p = 0.17$). In addition, the experts evaluated each kidney independently without regard to whether or not the other kidney needed furosemide or was obstructed. For clinical purposes, the two kidneys were thus treated as independent. Zhang (16) proposed a method that allows for multiple, correlated binary responses. With bagging (which is in itself already computationally intensive), the computing requirements of Zhang's proposed approach are magnified. More computationally efficient methods may be investigated and implemented.

CONCLUSION

CART with bootstrap aggregation were successfully applied in the baseline plus furosemide diuresis renography protocol to determine when obstruction can be excluded without the furosemide acquisition. To our knowledge, this is the first time CART analysis with bagging has been applied in nuclear medicine; moreover, our data show that algorithms based on bootstrap aggregation are likely to improve results and stabilize performance. CART analysis also provided the important insight that simple voiding parameters (ratios of the postvoid counts in the kidney to the maximum kidney counts and the 1- to 2-minute kidney counts) are likely to be more robust discriminators between obstruction and nonobstruction than the $T_{1/2}$. For physicians who lack extensive experience in diuresis renography, CART has the potential to serve as a decision support tool to help avoid the cost and patient inconvenience of unnecessary furosemide imaging. Finally, this initial analysis will serve as a platform for applying CART to more challenging problems in renal nuclear medicine such as obstruction and renovascular hypertension and this approach may also be adapted to address a broader range of imaging problems in diagnostic radiology.

REFERENCES

1. Morgan JN, Sonquist JA. Problems in the analysis of survey data and a proposal. *J Am Stat Assoc.* 1963; 58:415-434.

2. Morgan, JN.; Messenger, RC. THAID: A Sequential Search Program for the Analysis of Nominal Scale Dependent Variables. Survey Research Center, Institute for Social Research, University of Michigan; 1973.
3. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Classification and Regression Trees. Wadsworth; Belmont, CA: 1984.
4. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; New York: 2001. p. 246-250.
5. O'Reilly P, Aurell M, Britton K, Kletter K, Rosenthal L, Testa T. Consensus on diuresis renography for investigating the dilated upper urinary tract. J Nucl Med. 1996; 37:1872–1876. [PubMed: 8917195]
6. Kuyvenhoven J, Piepsz A, Ham H. When could the administration of furosemide be avoided? Clin Nucl Med. 2003; 28:732–737. [PubMed: 12972994]
7. Taylor A, Corrigan PL, Galt J, et al. Measuring technetium-99m-MAG3 clearance with an improved camera-based method. J Nucl Med. 1995; 36:1689–1695. [PubMed: 7658232]
8. Taylor A, Manatunga A, Morton K, et al. Multicenter trial validation of a camera based method to measure Tc-99m mercaptoacetyl triglycine (MAG3) clearance. Radiology. 1997; 204:47–54. [PubMed: 9205222]
9. Efron, B.; Tibshirani, R. An Introduction to the Bootstrap. Chapman and Hall; London: 1993.
10. Comprehensive R Archive Network (CRAN). Available at <http://www.r-project.org>
11. Venables, WN.; Ripley, BD. Modern Applied Statistics With S-Plus. 3rd ed.. Springer Verlag; New York: p. 303-310.200
12. Maindonald, J.; Braun, J. Data Analysis and Graphics Using R: An Example-Based Approach. Cambridge University Press; Cambridge: 2003. Chapter 10
13. Piepsz A, Kuyvenhoven JD, Tondeur M, Ham H. Normalized residual activity: Usual values and robustness of the method. J Nucl Med. 2002; 43:33–38. [PubMed: 11801700]
14. Thrall, JH.; Ziessman, HA. Nuclear Medicine: The Requisites. Mosby-Year Book; St Louis: 1995. p. 302
15. Diggle, PJ.; Heagerty, P.; Liang, KY.; Zeger, SL. Analysis of Longitudinal Data. 2nd ed.. Oxford University Press; Oxford: 2002.

Figure 1.

Original tree for predicting need for furosemide (yes/no) in the right kidney training set. ^aNumber of misclassified right kidneys in this terminal node. ^bNumber of right kidneys that fell in this terminal node.

Table 1

Distribution of misclassification rates for the validation samples

Kidney	n	Original tree	Bootstrapped trees ($k = 1001$)					Max	Bagging
			Mean	Median	SD	Min	Max		
Right	64	15.62	15.87	15.63	6.98	4.69	35.94	10.94	
Left	60	8.33	11.53	10.00	3.05	6.67	30.00	10.00	

n = number of kidneys

k = number of bootstrapped trees

Table 2

Number (percent) of bootstrapped trees on the training set containing one or more levels

Level	Right kidney	Left kidney
1	37 (3.7%)	92 (9.2%)
2	257 (25.7%)	416 (41.6%)
3	448 (44.8%)	377 (37.7%)
4	232 (23.1%)	105 (10.5%)
5	27 (2.7%)	10 (1.0%)
6	0 (0.0%)	1 (0.1%)

Table 3

Most frequent kidney parameters in the training set on level 1

Kidney parameter	Frequency	
	Right kidney	Left kidney
Cortical 20 min/max ratio	32.1%	3.6%
Postvoid/max ratio	29.1%	83.0%
Postvoid/1-2 min ratio	13.3%	6.3%
19-20 min/max ratio	10.8%	1.0%