DOI: 10.1111/j.1475-6773.2007.00761.x
METHODS ARTICLE

Regression Tree Boosting to Adjust Health Care Cost Predictions for Diagnostic Mix

John W. Robinson

Objective. To assess the ability of regression tree boosting to risk-adjust health care cost predictions, using diagnostic groups and demographic variables as inputs. Systems for risk-adjusting health care cost, described in the literature, have consistently employed deterministic models to account for interactions among diagnostic groups, simplifying their statistical representation, but sacrificing potentially useful information. An alternative is to use a statistical learning algorithm such as regression tree boosting that systematically searches the data for consequential interactions, which it automatically incorporates into a risk-adjustment model that is customized to the population under study.

Data Source. Administrative data for over 2 million enrollees in indemnity, preferred provider organization (PPO), and point-of-service (POS) plans from Thomson Medstat's Commercial Claims and Encounters database.

Study Design. The Agency for Healthcare Research and Quality's Clinical Classification Software (CCS) was used to sort 2001 diagnoses into 260 diagnosis categories (DCs). For each plan type (indemnity, PPO, and POS), boosted regression trees and main effects linear models were fitted to predict concurrent (2001) and prospective (2002) total health care cost per patient, given DCs and demographic variables.

Principal Findings. Regression tree boosting explained 49.7–52.1 percent of concurrent cost variance and 15.2–17.7 percent of prospective cost variance in independent test samples. Corresponding results for main effects linear models were 42.5–47.6 percent and 14.2–16.6 percent.

Conclusions. The combination of regression tree boosting and a diagnostic grouping scheme, such as CCS, represents a competitive alternative to risk-adjustment systems that use complex deterministic models to account for interactions among diagnostic groups.

Key Words. Risk adjustment, case mix, health care cost, boosting, data mining

Models that use diagnoses from claims to risk-adjust health care cost predictions are widely employed by health services researchers and public and private payers. In provider profiling applications, risk-adjustment models are

used to estimate an expected cost for each of a provider's patients, to be compared with each patient's observed cost. Comparisons of observed and expected costs are then aggregated across a provider's patient sample to yield an overall assessment of provider performance (Powe et al. 1996; Thomas, Grazier, and Ward 2004a; Robinson, Zeger, and Forrest 2006). In capitation-setting applications, risk-adjustment models are used to estimate an expected annual cost for each patient to be enrolled in a prepaid health plan. Expected costs are then summed to yield an expected annual cost for the entire enrollment (Fowles et al. 1996; Ash et al. 2000; Mark et al. 2003). Here *prediction* refers to model-based estimation of an observation's value, regardless of its timing.

Diagnoses are recorded on claims in the United States using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Owing to the vast number (over 14,000) of ICD-9-CM diagnoses (Iezzoni 2003), risk-adjustment systems generally begin by sorting them into a manageable number of mutually exclusive groups, based on similarity of clinical features and resource demands. For example, three risk-adjustment systems, Adjusted Clinical Groups (ACGs) (Weiner et al. 1991; Health Services Research and Development Center 2001), Diagnostic Cost Groups/Hierarchical Condition Categories (DCG/HCCs) (Ash et al. 2000), and Clinical Risk Groups (CRGs) (Hughes et al. 2004), sort the ICD-9-CM diagnoses into 32, 118, and 534 diagnostic groups, respectively.

The ACG, DCG/HCC, and CRG systems are proprietary. An alternative grouping scheme that is in the public domain is the Clinical Classification Software (CCS), developed and continually updated by the Agency for Healthcare Research and Quality (AHRQ) (Elixhauser, Steiner, and Palmer 2005). CCS sorts the ICD-9-CM diagnoses into 260 mutually exclusive diagnosis categories (DCs) based on clinical similarity. The DCs have been used for risk adjustment by representing each as a dummy variable onto which cost is regressed (Cowen et al. 1998), making the implicit assumption that each DC's effect on cost is independent of the presence of any other DC. This amounts to adjusting for the main effects of DCs but not for interactions among them (Searle 1971).

Address correspondence to John W. Robinson, M.D., Ph.D., LLC, Healthcare Management and Statistical Consulting, 4303 Stanford Street, Chevy Chase, MD 20815; e-mail: johnwrobinson@verizon.net.

INTERACTIONS AMONG DIAGNOSTIC GROUPS

Interactions correspond clinically to the effects of comorbidity and complications. Positive interaction occurs when two diagnoses are more costly to manage together in one individual than separately in two. For example, a respiratory tract infection and chronic obstructive pulmonary disease occurring together in the same person would be much more likely to require hospital admission than either condition alone. Negative interaction occurs when two conditions are less costly to treat together in one individual than separately in two, such as when the same procedure is needed to manage each of two pathologically independent conditions. If a patient's diagnoses belong to more than one diagnostic group, analogous interactions can occur at the diagnostic group level.

The number of potential interactions among diagnostic groups is vast, even if only low-order interactions are considered. For example, restricting attention to interactions involving six or fewer diagnostic groups still yields over 500 million potential interactions among the 32 diagnostic groups in the ACG system. However, even if it were feasible to represent every low-order interaction by a model parameter, it would not be desirable to do so, because only a small fraction of interactions are likely to be consequential in any particular application. Thus, an approach is needed for selecting consequential interactions to incorporate into a risk-adjustment model. In general, two alternatives are possible: (1) create a deterministic model that explicitly or implicitly anticipates the magnitude and direction of every possible interaction in any application, or (2) employ a statistical learning algorithm that systematically explores the data at hand, finds consequential interactions, and automatically incorporates them into a risk-adjustment model (Breiman 2001; Hastie, Tibshirani, and Friedman 2001).

The first approach has been consistently employed by risk-adjustment systems described in the literature. For example, under the DCG/HCC system, cost is regressed on dummy variables for the 118 diagnostic groups, termed "condition categories," to estimate a main-effects risk-adjustment model. However, before assigning values to the 118 dummy variables, each person's set of condition categories is modified using a system of hierarchical rules that has the effect of accounting for selected potential negative interactions. For example, if a person with metastatic cancer has diagnoses in condition categories representing both the primary cancer and metastatic spread, only the condition category representing metastatic spread is assigned to that person (Ash et al. 2000).

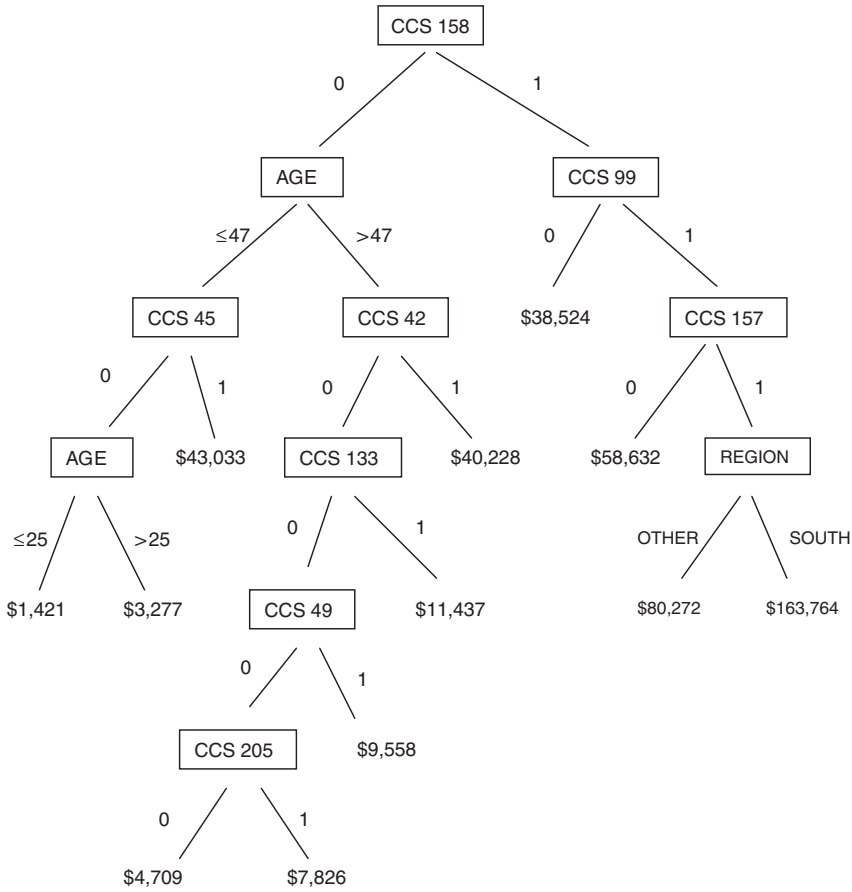
Under the ACG and CRG systems, clinical reasoning and historical data analysis have been used to map every possible combination of diagnostic groups, age, and gender to one of a collection of mutually exclusive risk categories: the 93 ACGs or 1,075 CRGs. Thus, each risk category represents potentially millions of different combinations of diagnostic groups, all of which are assumed to share the same effect on cost. Estimates of cost effects of risk categories are obtained by regressing cost on dummy variables representing the categories (Weiner et al. 1991; Health Services Research and Development Center 2001; Hughes et al. 2004). The above is not meant to imply that the only purpose of ordering diagnostic groups into hierarchies or assigning them to risk categories is to handle potential interactions. In some instances, these strategies have been used to thwart the anticipated effects of redundant diagnoses on cost predictions (Ash et al. 2000).

Regression Tree Boosting

The examples just described use deduction and historical data analysis to create a deterministic model of interactions among diagnostic groups. However, it might not be possible to devise a deterministic model that adequately represents such a large and complex system and is sufficiently flexible to adapt to different populations and conditions. An alternative approach is to forgo attempting to forecast the magnitude and direction of interactions a priori, and instead use a statistical learning algorithm that systematically explores the data at hand, finds consequential interactions, and automatically incorporates them into a risk-adjustment model, which is thus customized to the population and conditions under study.

Regression tree boosting, which involves the iterative fitting of many small regression trees, is a statistical learning method that is especially useful for prediction of a continuous variable, such as cost, based on the values of a very large number of potentially interacting categorical and continuous variables (Friedman 2001; Hastie, Tibshirani, and Friedman 2001). Figure 1 shows the first of many regression trees from a boosting sequence fit to predict total health care cost given demographic variables and prior-year DCs. The tree contains 11 internal nodes, each associated with a splitting variable and split point, and 12 terminal nodes, each associated with a dollar-valued coefficient. The tree divides the sample into 12 disjoint subsamples, each defined by a unique combination of splitting variable values, and assigns a coefficient to each. The coefficients incorporate both main effects of splitting variables and

Figure 1: Regression Tree Example.



First of 232 regression trees from final boosted regression trees model for 2002 cost prediction among preferred provider organization members. Boxes represent internal nodes and contain names of splitting variables. Lines projecting below boxes are accompanied by splitting variable values that fall on either side of split point. Dollar figures correspond to terminal node coefficients. Clinical Classification Software diagnosis categories (DCs), coded 1 if present and 0 if absent, are as follows: CCS 158 = chronic renal failure; CCS 99 = hypertension with complications and secondary hypertension; CCS 45 = maintenance chemotherapy and radiotherapy; CCS 42 = secondary malignancies; CCS 157 = acute and unspecified renal failure; CCS 133 = other lower respiratory disease; CCS 49 = diabetes mellitus without complication; CCS 205 = spondylosis, intervertebral disk disorders, and other back problems.

interactions among them. The number of internal nodes determines the highest possible order of interaction that a coefficient can represent. Because this is the first tree, each coefficient represents the predicted cost for all observations in the corresponding subsample.

The tree in Figure 1 was "grown" using a *best-first* strategy (Friedman, Hastie, and Tibshirani 2000), whereby internal node splits are accomplished one at a time. At each step, for each currently terminal node, the splitting variable and split point are identified that would achieve the greatest improvement in overall fit between tree-based predictions and observed values. The potential improvements are compared, and a split is carried out at the currently terminal node whose split results in the greatest improvement in fit. To lessen the potential influence of any extreme outlier, a minimum terminal node size is set, for example at 10 observations, thus restricting eligible splits to those that yield at least this minimum number of observations on either side. The process ceases when a predetermined number of terminal nodes is reached. The internal node splitting variables and split points and the terminal node coefficients are the estimated parameters that define the fitted tree.

For tree boosting, the predicted costs due to fitting the first tree are subtracted from corresponding observed costs and a second tree is fitted to the resultant residuals. The second tree, which generally splits on different variables than the first, divides the sample into 12 disjoint subsamples and assigns a coefficient to each. To "boost" the cost prediction for each observation, a fraction—termed the *scaling fraction*, with a typical value of 0.1—of the appropriate coefficient from the second tree is added to the prediction from the first tree. The resultant boosted predictions are, on average, closer to the observed costs than predictions based on only the first tree. The boosted predictions are then subtracted from observed costs, yielding new residuals, to which a third tree is fitted, and so on.

This process is repeated until additional trees no longer improve the fit between observed and predicted costs in a *validation* sample that has been randomly drawn from the same population as the *training* sample that has been used to estimate the tree parameters (Friedman 2001; Hastie, Tibshirani, and Friedman 2001). The reason for using an independent validation sample to assess fit is to avoid selecting a final set of *tuning parameters* (e.g., number of terminal nodes per tree, scaling fraction, minimum number of observations per terminal node) that predicts well in the training sample, but not in a fresh sample from the same population. A final boosted trees model comprises additive contributions from typically hundreds of small trees and, thus, can be represented as an additive model (Friedman 2001; Hastie, Tibshirani, and Friedman 2001).

Formed in this manner, a final boosted trees model is intended to give an accurate and precise cost prediction but not intended to describe the true mechanism that generates cost from diagnostic mix. However, an interpretation of the influence on cost of each of the diagnostic groups and other independent variables is provided by relative importance statistics, which are described in "Methods."

I have been unable to find a published use of regression tree boosting to risk-adjust health care cost predictions using diagnostic data. Regression tree boosting has been successfully used to predict the cost of inpatient rehabilitation given age, type of impairment, and continuous measures of motor and cognitive functioning (Relles, Ridgeway, and Carter 2002), and a related technique, classification tree boosting, has been shown to predict mortality in intensive care more accurately than logistic regression given age, gender, and 12 clinical measures (Neumann et al. 2004). Both of these applications involved far fewer independent variables than are needed to represent the 260 DCs.

The principal aim of this study is to assess the ability of regression tree boosting to risk-adjust concurrent and prospective health care cost predictions using diagnostic groups and demographic variables as inputs. A secondary aim is to assess the ability of regression tree boosting to identify consequential interactions among diagnostic groups and incorporate them into a risk-adjustment model. This latter ability can be measured by comparing the fit of boosted regression trees models with main effects linear models, which do not account for interactions.

METHODS

Data Source and Variables

To demonstrate the use of regression tree boosting, I employ Thomson Medstat's MarketScan Commercial Claims and Encounters Database (2004), which includes demographic information and claims histories for all persons enrolled in selected health plans of 45 private and public employers. The database includes specific information about each enrollee's benefits, but does not name the employers or health plans or provide enrollees' insurance identification numbers. All 50 states and the District of Columbia are represented.

Dependent variables are total paid costs for 2001 and 2002, including inpatient, outpatient, and pharmacy costs. Independent variables for predicting 2001 total cost are gender, age, 2001 diagnoses, months enrolled in 2001, and

2001 region of residence, and for predicting 2002 total cost are gender, age, 2001 diagnoses, months enrolled in 2002, and 2002 region of residence. ("Months enrolled" refers to enrollment duration, measured in number of months.)

I included in the study sample only persons who were enrolled in a health plan for at least 1 month of both 2001 and 2002, so that both years' predictions would be based on the same mix of 2001 diagnoses. I included only persons who were enrolled in plans that had a pharmacy benefit and did not capitate payments to providers, so that total costs would be available. Because the database does not include Medicare claims, I included only persons who would turn 65 after 2002. The Commercial Claims and Encounters database contains 2,758,476 persons meeting these inclusion criteria. Of these, 2,320,043 (84.1 percent) had a claim in 2001 and were thus eligible for inclusion in the study sample. Persons without a 2001 claim were not included, because doing so would have meant using the absence of a 2001 claim to predict 2001 cost, a circular function that would have artificially inflated the precision of 2001 cost predictions.

All 2001 claims for services involving face-to-face encounters with health care practitioners were used to create each person's list of diagnoses. Then, using each list and the CCS (Elixhauser, Steiner, and Palmer 2005), values were assigned to dummy variables representing the 260 DCs.

The dependent variable, total annual paid cost, representing the sum of plan and enrollee liabilities, incorporates contractual discounts applied by health plans to providers' fees. Because the fraction of claims involving such discounts likely differs by type of plan—for example, indemnity plans generally employ less discounting than preferred provider organization (PPO) and point-of-service (POS) plans (Dudley and Luft 2001)—paid costs represent different quantities under different types of plan. Hence, I fit separate models for each type of plan that was well represented in the study sample.

Boosting Algorithms

I wrote programs in SAS Interactive Matrix Language (SAS Institute 2004a) that implement the "LS_Boost" and "*M*-Tree Boost" algorithms outlined by Friedman (2001). The algorithms both use a best-first strategy for fitting each tree, but use different criteria to optimize tree fit. LS_Boost minimizes the sum of squared residuals, the familiar criterion used to fit linear models. However, the sum of squared residuals can be excessively influenced by observations with unusually large residuals (Huber 1980). To constrain the influence of

extreme residuals, *M*-Tree Boost uses *Huber loss* as its fit criterion, whereby the contribution of the very largest residuals is set proportional to their absolute rather than squared values (Friedman 2001). The fraction of residuals that have their influence constrained in this manner, referred to here as the *breakdown fraction*, is a tuning parameter of the boosting algorithm.

I found that *M*-Tree Boost with a breakdown fraction of 0.0001 consistently resulted in a better validation sample fit than LS_Boost, thus all reported results are based on *M*-Tree Boost with a breakdown fraction of 0.0001, meaning that the influence of the 1 in 10,000 largest residual costs was constrained in the fitting of each sequential tree. (Larger breakdown fractions improved the precision of cost prediction, but introduced a significant negative bias.)

For each of the six combinations of health plan type (indemnity, PPO, and POS) and dependent variable (2001 and 2002 total cost) for which models were to be estimated, I randomly divided the relevant portion of the study sample into *training*, *validation*, and *test* samples, containing 50, 25, and 25 percent of observations, respectively (Hastie, Tibshirani, and Friedman 2001). Trees were fitted to training sample observations only. During each run of a boosting algorithm, observations in the validation sample were "run-down" successively fitted trees, meaning that the internal node parameters from each tree fitted to the training sample were used to divide the validation sample into disjoint subsamples to which the terminal coefficients from the fitted tree were applied, thus continually updating predicted values in the validation sample. When additional trees no longer improved the overall fit between observed and predicted values in the validation sample, the algorithm was stopped.

The validation samples were also used to select a best-fitting final boosted regression trees model for each plan type and cost year, defined by values of the *tuning parameters*: number of terminal nodes per tree, scaling fraction, minimum number of observations per terminal node, loss criterion (sum of squared residuals or Huber), and breakdown fraction (for Huber loss). Typically, five to 10 boosting runs were needed to identify optimal values of the tuning parameters for a combination of plan type and cost year. Once a final model had been selected, the relevant test sample was "run-down" that model's tree sequence, yielding an assessment of the model's predictive ability that could be generalized to an independent sample from the same population as the training sample. (The validation sample could not have been used for this purpose, because it had been used to select a final model [Hastie, Tibshirani, and Friedman 2001].)

Model Assessment

As a measure of the predictive ability of regression tree boosting, let the *percent of variance explained* by tree boosting be defined as {1–[(sum of squared residuals from a boosted regression trees model)/(sum of squared residuals from an intercept-only model)]} × 100 percent, where "residuals" refers to test sample residuals. This quantity is conceptually equivalent to the multiple correlation coefficient, R^2, which it would equal were the boosted trees model replaced with a linear model and the training sample reused as the test sample (Searle 1971). Under an intercept-only model, the residual for each test sample observation is the difference between its observed value and the mean cost in the corresponding training sample.

For benchmarks against which to compare the predictive ability of tree boosting, I fit main effects linear models to each of the six training samples, by regressing cost on dummy variables for the 260 DCs, gender, age (classified in decades), region of residence, and months enrolled (classified as 1–3, 4–6, 7–9, or 10–12 months) using the SAS GLM procedure (SAS Institute 2004b). Each of the six main effects linear models was used to compute a percent of variance explained in the appropriate test sample, for comparison with the percent of variance explained by tree boosting.

For each final boosted trees model, the relative importance of each independent variable to model fit was measured as suggested by Friedman (2001). First, for each independent variable, the reduction in Huber loss was summed across all internal nodes, of all trees, that split on that variable and divided by the total number of internal nodes (number of internal nodes per tree × number of trees), yielding a squared *importance* for that variable. Once a squared importance had been obtained for each variable, the square root of the largest squared importance was divided into the square root of each other squared importance to obtain a *relative importance* for each independent variable, in the range of 0–100 percent.

RESULTS

Sample Characteristics

Table 1 presents sample characteristics stratified by 2001 plan enrollment. Of the 2,320,043 persons in the sample, 30.4, 47.9, and 19.6 percent were enrolled in indemnity, PPO, and POS plans, respectively, and the remaining 2.1 percent changed plans during the year. More than 20 percent of patients in each plan type had diagnoses in six or more DCs, and 9 percent of the total

Table 1: Sample Characteristics by 2001 Plan Enrollment (N = 2,320,043)

	Plan Type			
	Indemnity (n = 704,198)	*PPO (n = 1,111,690)*	*POS (n = 454,237)*	*Mixed (n = 49,918)*
Total annual cost (dollars)				
Mean	3,972	3,163	2,823	3,186
Median	1,258	993	834	986
90th percentile	8,874	7,079	6,227	7,248
99th percentile	42,577	32,642	29,901	33,025
Characteristic (%)				
Gender				
Female	55.6	54.5	57.2	54.5
Age				
0–17	17.9	22.8	25.1	23.9
18–39	21.0	24.1	31.2	25.6
40–49	18.4	20.0	21.4	17.6
50–63	42.7	33.0	22.3	32.9
Region				
Northeast	4.0	10.5	9.1	2.4
North Central	70.0	29.0	25.3	10.3
South	22.8	43.8	51.3	39.6
West	3.0	16.6	14.3	47.4
Unknown	0.1	0.1	0.0	0.2
Months enrolled				
1–3	1.9	2.4	2.4	1.6
4–6	2.0	3.4	2.8	1.5
7–9	1.7	2.7	3.8	3.0
10–12	94.4	91.5	91.0	93.9
No. of DCs*				
0	12.5	7.0	7.5	6.8
1–2	34.3	35.6	35.1	34.3
3–5	30.7	33.9	34.3	35.3
6–10	17.6	18.7	18.7	19.0
11+	4.9	4.8	4.4	4.6

*Number of Clinical Classification Software diagnosis categories (DCs).

POS, point-of-service; PPO, preferred provider organization; "Mixed" includes persons who changed plans during 2001.

sample had no diagnosis assigned, due to having a claim but no face-to-face clinical encounter (e.g., claims for laboratory services only).

Between 2001 and 2002, 12 percent of the sample changed plans, resulting in decreases in indemnity and POS enrollment, an increase in PPO enrollment, and enrollment in exclusive provider organizations (EPOs), a plan

type not represented in 2001. In 2002, 27.5, 52.4, 14.8, and 4.2 percent of the sample were enrolled in indemnity, PPO, POS, and EPO plans, respectively, and 1.1 percent changed plans. Mean 2002 costs were $4,188, $3,907, $2,969, and $3,474 in indemnity, PPO, POS, and EPO plans, respectively. Distributions of other sample characteristics in 2002 were similar to 2001.

Predictive Performance

I fit boosted regression trees and main effects linear models to the six subsamples enrolled in indemnity, PPO, and POS plans in 2001 and 2002. Persons who had changed plans during a given year were excluded from that year's analysis, and EPO enrollees were excluded from the 2002 analysis. Based on validation sample fit, a final boosted trees model was selected for each combination of plan type and cost year. Results for the final boosted regression trees and main effects linear models are shown in Table 2.

As noted in "Methods," Huber loss with a breakdown fraction of 0.0001 consistently yielded the best validation sample fit, as did specifying a minimum of 10 observations per terminal node and a scaling fraction of 0.1. Thus, the only tuning parameter that differed among the final boosted trees models was the number of terminal nodes per tree. Values for this tuning parameter are included in Table 2.

Table 2: Final Boosted Regression Trees and Main Effects Linear Models

			Percent of Variance Explained†			*Mean Cost (Dollars)*†		
							Predicted	
Year	*Plan Type*	*Nodes/Tree**	*Boosted Trees*	*Main Effects*	*Gain*‡	*Observed*	*Boosted Trees*	*Main Effects*
2001	Indemnity	9	49.7	47.2	2.5	4,005	3,986	4,007
	PPO	9	52.1	47.6	4.5	3,141	3,136	3,157
	POS	6	49.8	42.5	7.3	2,787	2,783	2,805
2002	Indemnity	8	17.7	16.6	1.1	4,141	4,184	4,216
	PPO	12	15.2	14.2	1.0	3,898	3,885	3,910
	POS	2	15.2	14.8	0.4	2,953	2,982	2,994

*Number of terminal nodes per tree in final boosted regression trees model.

†Based on test samples.

‡Gain = percent of variance explained by tree boosting minus percent explained by main effects.

POS, point-of-service; PPO, preferred provider organization.

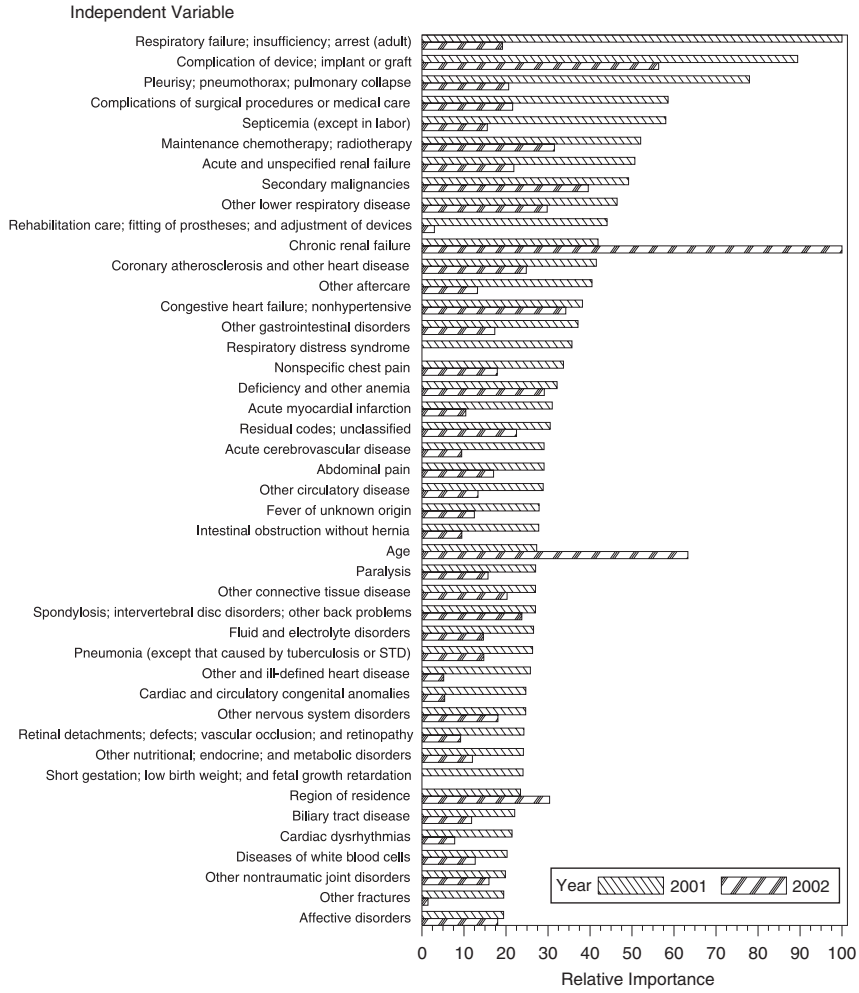
Table 2 shows that the percent of variance explained by tree boosting ranges from 49.7 to 52.1 for concurrent risk adjustment and from 15.2 to 17.7 for prospective risk adjustment. These quantities consistently exceed the corresponding percentages for main effects linear models, by 2.5–7.3 percent for concurrent prediction and by 0.4–1.1 percent for prospective prediction. Table 2 also shows that mean predicted costs in the test samples based on boosted trees and linear models are similarly unbiased for mean observed costs. Thus, the gain in precision due to tree boosting is not accompanied by any loss of overall accuracy.

The number of terminal nodes per tree ranges from 6 to 12 for all but one of the final boosted trees models; thus interactions involving as many as five to 11 DCs are able to contribute to cost prediction in five of six final models. The number of additive trees in a final model, which is inversely related to the number of nodes per tree, ranges from 232 in the 2002 PPO model, with 12 terminal nodes, to 798 in the 2001 POS model, with six terminal nodes, indicating that a very large number of interactions among DCs contribute to the fit of these five models. (The numbers of trees are 296, 335, and 624 in the 2002 indemnity, 2001 PPO, and 2001 indemnity final models, respectively.) The sixth final model, for prospective prediction among POS plan members, comprises 1,644 trees, each with one internal node (and two terminal nodes), thus incorporating no interactions. Not surprisingly, this model shows the smallest gain in percent of variance explained over the corresponding main effects linear model.

Relative Importance of Independent Variables

Figure 2 displays the relative importance of independent variables for 2001 and 2002 risk adjustment, based on the final boosted regression trees models for PPO members, ordered by 2001 relative importance. Only the 44 independent variables for which 2001 relative importance is largest are included in Figure 2. (Relative importance of all 264 variables can be found in "Supplementary Material.") Of greatest importance for concurrent (2001) risk adjustment are diagnosis groups representing very serious conditions, many of which might involve admission to intensive care. Most of these same conditions are important for prospective (2002) risk adjustment, but to a lesser extent. Acute conditions tend to be more important for concurrent risk adjustment, whereas chronic conditions tend to be more important for prospective risk adjustment, as exemplified by comparison of relative importance patterns for acute and chronic renal failure. Analogous importance charts

Figure 2: Relative Importance of Independent Variables for Predicting 2001 and 2002 Total Health Care Cost for Preferred Provider Organization Enrollees Based on Final Boosted Regression Trees Models.



Independent variables are listed in order of importance for predicting 2001 cost. Only the 44 variables for which 2001 relative importance is largest are shown. Region of residence is for the corresponding year. (A figure displaying relative importance of all 264 independent variables can be found in "Supplementary Material.")

based on final boosted trees models for indemnity and POS samples demonstrate similar findings (and are available in "Supplementary Material").

DISCUSSION

Regression tree boosting, using CCS DCs and demographic variables as inputs, explained 50–52 percent of the variance in concurrent annual health care cost and 15–18 percent of the variance in prospective annual cost. Comparison of these results with results reported using other risk-adjustment methods must be undertaken cautiously, because the percent of variance explained by any risk-adjustment procedure depends partly on the population cost distribution, which differs between studies. Nevertheless, the percent of variance explained by regression tree boosting appears to equal or exceed results reported using deterministic models that rely on diagnostic groups, age, and gender as inputs (Weiner et al. 1991; Fowles et al. 1996; Ash et al. 2000; Hughes et al. 2004; Thomas, Grazier, and Ward 2004b).

Regression tree boosting appears to be a very effective means of finding consequential interactions among diagnostic groups and incorporating them into risk adjustment, as evidenced by its substantial advantage over main effects linear models, which do not account for interactions. The advantage of boosted regression trees over main effects linear models is more pronounced when predicting cost concurrently than prospectively, evidently because interactions among diagnostic groups are more predictive of concurrent than prospective cost.

A small portion of the advantage of boosted regression trees over main effects linear models is apparently due to features unrelated to representing interactions. The best-fitting 2002 POS model, with just one internal node per tree, does not model interactions, yet explains slightly more variance than the corresponding main effects linear model. Features of the boosting algorithm that might explain this advantage include the use of Huber loss, which constrains the influence of extremely high-cost observations (without trimming them), and fractional scaling, which has an effect akin to parameter shrinkage (Hastie, Tibshirani, and Friedman 2001). Both of these features resist overfitting the training sample.

The results reported here suggest that regression tree boosting may obviate the need to process diagnostic groups through a deterministic interaction model before using them for risk adjustment. Deterministic interaction models are labor intensive to develop, involving teams of health

services researchers and clinical specialists (Weiner et al. 1991; Ash et al. 2000; Hughes et al. 2004), and incorporate scores of assumptions and restrictions that limit their adaptability to different populations and conditions. On the other hand, absent any prior assumption about the direction or magnitude of any potential interaction, tree-boosting algorithms effectively identified consequential interactions and automatically incorporated them into risk-adjustment models. As a result, risk adjustment was accomplished very efficiently and flexibly, using a publicly available diagnosis classifier and an algorithm implemented by a single researcher.

REFERENCES

Ash, A. S., R. P. Ellis, G. C. Pope, J. Z. Ayanian, D. W. Bates, H. Burstin, L. I. Iezzoni, E. MacKay, and W. Yu. 2000. "Using Diagnoses to Describe Populations and Predict Costs." *Health Care Financing Review* 21 (3): 7–28.

Breiman, L. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–231.

Cowen, M. E., D. J. Dusseau, B. G. Toth, C. Guisinger, M. W. Zodet, and Y. Shyr. 1998. "Casemix Adjustment of Managed Care Claims Data Using the Clinical Classification for Health Policy Research Method." *Medical Care* 36 (7): 1108–13.

Dudley, R. A., and H. S. Luft. 2001. "Managed Care in Transition." *New England Journal of Medicine* 344 (14): 1087–91.

Elixhauser, A., C. Steiner, and L. Palmer. 2005. "Clinical Classifications Software (CCS). U.S. Agency for Healthcare Research and Quality" [accessed on October 20, 2005]. Available at http://www.ahrq.gov/data/hcup/ccs.htm#download

Fowles, J. B., J. P. Weiner, D. Knutson, E. Fowler, A. M. Tucker, and M. Ireland. 1996. "Taking Health Status into Account When Setting Capitation Rates: A Comparison of Risk-Adjustment Methods." *Journal of the American Medical Association* 276 (16): 1316–21.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–232.

Friedman, J., T. Hastie, and R. Tibshirani. 2000. "Additive Logistic Regression: A Statistical View of Boosting." *Annals of Statistics* 28 (2): 337–407.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning.* New York: Springer.

Health Services Research and Development Center at The Johns Hopkins University Bloomberg School of Public Health. 2001. *The Johns Hopkins ACG Case-Mix System, Version 5.0.* Baltimore: The Johns Hopkins University.

Huber, P. 1980. *Robust Statistics.* New York: John Wiley & Sons.

Hughes, J. S., R. F. Averill, J. Eisenhandler, N. I. Goldfield, J. Muldoon, J. M. Neff, and J. C. Gay. 2004. "Clinical Risk Groups (CRGs): A Classification System for Risk-Adjusted Capitation-Based Payment and Health Care Management." *Medical Care* 42 (1): 81–90.

Iezzoni, L. I. 2003. "Coded Data from Administrative Sources." In *Risk Adjustment for Measuring Health Care Outcomes*, edited by L. I. Iezzoni, pp. 83–138. Chicago: Health Administration Press.

Mark, T. L., R. J. Ozminkowski, A. Kirk, S. L. Ettner, and J. Drabek. 2003. "Risk Adjustment for People with Chronic Conditions in Private Sector Health Plans." *Medical Decision Making* 23: 397–405.

Neumann, A., J. Holstein, J-R. Le Gall, and E. Lepage. 2004. "Measuring Performance in Health Care: Case-Mix Adjustment by Boosted Decision Trees." *Artificial Intelligence in Medicine* 32: 97–113.

Powe, N. R., J. P. Weiner, B. Starfield, M. Stuart, A. Baker, and D. M. Steinwachs. 1996. "Systemwide Provider Performance in a Medicaid Program: Profiling the Care of Patients with Chronic Illnesses." *Medical Care* 34 (8): 798–810.

Relles, D., G. Ridgeway, and G. Carter. 2002. "Data Mining and the Implementation of a Prospective Payment System for Inpatient Rehabilitation." *Health Services Research and Outcomes Methodology* 3: 247–66.

Robinson, J. W., S. L. Zeger, and C. B. Forrest. 2006. "A Hierarchical Multivariate Two-Part Model for Profiling Providers' Effects on Health Care Charges." *Journal of the American Statistical Association* 101: 911–23.

SAS Institute. 2004a. *SAS/IML 9.1 User's Guide.* Cary, NC: SAS Institute Inc.

———. 2004b. *SAS/STAT 9.1 User's Guide.* Cary, NC: SAS Institute Inc.

Searle, S. R. 1971. *Linear Models.* New York: John Wiley & Sons.

Thomas, J. W., K. L. Grazier, and K. Ward. 2004a. "Economic Profiling of Primary Care Physicians: Consistency among Risk-Adjusted Measures." *Health Services Research* 39 (4): 985–1003.

———. 2004b. "Comparing Accuracy of Risk-Adjustment Methodologies Used in Economic Profiling of Physicians." *Inquiry* 41: 218–31.

Thomson Medstat. 2004. *MarketScan Commercial Claims and Encounters Research Database.* Ann Arbor, MI: Thomson Medstat.

Weiner, J. P., B. H. Starfield, D. M. Steinwachs, and L. M. Mumford. 1991. "Development and Application of a Population-Oriented Measure of Ambulatory Care Case-Mix." *Medical Care* 29 (5): 452–72.

SUPPLEMENTARY MATERIAL

The following supplementary material for this article is available:

Figure A1. Relative Importance of Independent Variables for Predicting 2001 and 2002 Total Health Care Cost for **Preferred Provider Organization** Enrollees Based on Final Boosted Regression Trees Models.

Figure A2. Relative Importance of Independent Variables for Predicting 2001 and 2002 Total Health Care Cost for **Indemnity Plan** Enrollees Based on Final Boosted Regression Trees Models.

Figure A3. Relative Importance of Independent Variables for Predicting 2001 and 2002 Total Health Care Cost for **Point-of-Service Plan** Enrollees Based on Final Boosted Regression Trees Models.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1475-6773.2007.00761.x (this link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.