

Prognostic Groups in Colorectal Carcinoma Patients Based on Tumor Cell Proliferation and Classification and Regression Tree (CART) Survival Analysis

Vladimir A. Valera, MD, PhD,¹ Beatriz A. Walter, MD, PhD,^{2,3}
Naoyuki Yokoyama, MD, PhD,¹ Yu Koyama, MD, PhD,¹
Tsuneo Iiai, MD, PhD,¹ Haruhiko Okamoto, MD, PhD,¹
and Katsuyoshi Hatakeyama, MD, PhD, FACS¹

¹First Department of Surgery, Division of Digestive and General Surgery, Niigata University Graduate School of Medical and Dental Sciences, 1-757 Asahimachi Dori, Niigata City 951-8510, Japan

²Department of Otolaryngology, Niigata University Graduate School of Medical and Dental Sciences, Niigata City, Japan

³Third Department of Anatomy, Division of Microscopic Anatomy and Bioimaging, Niigata University Graduate School of Medical and Dental Sciences, Niigata City, Japan

Background: In this study, an alternative analytical method was used to model colorectal cancer (CRC) patients' long-term survival by assessing the prognostic value of the Ki-67 protein as a marker of tumor cell proliferation, and to illustrate the interaction between standard clinicopathologic variables and the proliferation marker in relation to their impact on survival.

Methods: A cohort of 106 surgically treated CRC patients was used for analysis. The expression of the cell-cycle-related Ki-67 protein in tumor samples was evaluated by immunohistochemistry. A score was assigned as the percentage of positive tumor cell staining, denoted as proliferation index (PI), and was used in a multivariate analysis using a recursive partitioning algorithm referred to as classification and regression tree (CART) to characterize the long-term survival after surgery.

Results: Of the covariates selected for their prognostic value, PI contributed most to the classification of survival status of patients. However, CART analysis selected the presence of distant metastasis as the best first split-up factor for predicting 5-year survival. CART then selected the following covariates for building up subgroups at risk for death: (1) PI; (2) pathological lymph node metastasis; (3) tumor size. Seven terminal subgroups were formed, with an overall misclassification rate of 16%.

Conclusions: These analyses demonstrated that a Ki-67-protein-based tumor proliferation index appeared as an independent prognostic variable that was consistently applied by the CART algorithm to classify patients into groups with similar clinical features and survival.

Key Words: Colorectal carcinomas—Proliferation index—Ki-67—CART analysis—Prognosis.

Current classification of colorectal carcinomas (CRC) for providing prognostic information is based

Received May 17, 2006; accepted June 2, 2006; published online October 7, 2006.

Address correspondence and reprint requests to: Vladimir A. Valera, MD, PhD; E-mail: vvalerar@med.niigata-u.ac.jp

Published by Springer Science+Business Media, Inc. © 2006 The Society of Surgical Oncology, Inc.

on the formal TNM staging system, which depends almost exclusively on the anatomical extent of disease and is assessed using a combination of tumor size or depth of invasion (T), lymph node spread (N), and presence or absence of metastases (M).¹ The clinical features and survival times of patients however can be sometimes highly heterogeneous for cases with cancers apparently TNM equivalent, making prog-

nostication at the individual level a more difficult task.²

The ability to predict tumor behavior based on factors other than tumor extent, identifying curable cancers or advanced-stage tumors with more favourable course, may significantly influence the selection of patients for additional therapy, potentially modifying pre or postoperative treatment plans, and finally may help to improve CRC-specific survival.³ However, the factors and the optimal strategies for selecting those patients have yet to be defined.

This has led to the massive current research on gene and protein expression analysis for the identification of possible prognostic factors. Individual molecular markers and patterns of markers are successfully subdividing traditional tumor classes into subsets that behave differently from each other.⁴ In this context, cell proliferation, as a result of several molecular pathways occurring in cancer cells, has been shown to correlate with tumor progression and patients' survival in several human solid neoplasms, as estimated by the expression of a cell-cycle related product, the protein Ki-67.⁵

Common survival analyses applied in those prognostic studies usually employ a form of logistic regression for the estimation of the risk for death after surgery, including the Cox proportional hazards model. However, the assumptions of the method regarding the linearity of the relationship between covariates and outcome are not always met, and the visualization and clinical use of the results are sometimes difficult to interpret by clinicians.⁶ An alternative method that overcome these issues is known as classification and regression tree (CART) analysis. CART analysis belongs to a family of nonparametric regression methods based on recursive partitioning of data that builds a decision tree structure and classifies subjects, finally generating groups of patients with similar clinical features and survival times.⁷ It can be used to explore the data, identify possible high-risk subgroups, and uncover interactions or effect changes among prognostic factors. Unlike logistic regression, CART analysis does not assume a multiplicative risk model or a specific parametric probability, does not require a specification of the risk function, and is not affected by outlying observations. Most importantly, the results of CART analysis are presented as a decision tree, which is intuitive, easier to understand than the results of many other statistical methods, and facilitates the visualization and interpretation of the interactions between factors related to survival, with a flow-chart form to find group allocation.^{8,9}

In this study, we applied this statistical method to model CRC patients' data in order to evaluate the prognostic capabilities of the cell-proliferation-related Ki-67 protein, to explore the interactions between tumor cell proliferation and other clinicopathologic variables and their impact on survival and to illustrate explicitly how these covariates interact.

METHODS

Patients' Tissue Samples

The cohort analyzed in this study is based on 106 CRC patients initially treated by surgery at Niigata University Hospital between 1 January 1995 and 31 December 1997. The medical records of these patients were reviewed and pathological diagnoses were entered into a database after excluding from further analysis those who (a) received preoperative chemo or radiotherapy; (b) had another type of cancer; or (c) had an associated colorectal disease. Of 188 patients with a diagnosis of CRC, 82 were excluded from further analysis based on the criteria outlined above. Clinical variables were analyzed within the following general categories: demographic variables, pathological variables, tumor burden, and involvement of specific metastatic sites. Eighty-five percent of the resections were performed with curative intent and patients with metastatic disease received standard fluorouracil-based postoperative adjuvant therapy based on 5-FU/LV for six cycles. This sample size provides 80% power at the 0.05 significance level to detect at least a 0.5 difference in hazard ratio for death between levels of any dichotomous covariate in a Cox multivariate model, with an estimated event probability of 30%.

Immunohistochemistry

Formalin-fixed, paraffin-embedded tissue blocks obtained from pathology archives were used for immunohistochemical staining. The procedure consisted in a double staining sequence, initially for the cell-cycle related protein Ki-67 followed by staining with an anticytokeratin antibody to differentiate proliferating epithelial cells from proliferating non-epithelial cells, as described previously.⁵ Only homogeneous brown nuclear staining for Ki-67 was considered positive. Negative controls resulted always in negative staining.

Assessment of proliferation index

Tumor samples were examined by two investigators without knowledge of the clinical or pathological

data. A minimum of two double-immunostained sections per patient were evaluated microscopically at a total magnification of 400 \times . A proliferation index (PI) was calculated for each sample as the percentage of Ki-67-positive tumor cell nuclei in relation to all anticytokeratin-stained epithelial cells in at least three high power fields including 1,000 tumor cells (range: 1,002–1,748), and was expressed as a mean value. In all cases, interobserver agreement as a measure of method reproducibility was 92% (intraclass correlation coefficient = 0.92).

Statistical Analysis

CART was implemented using 11 common clinical and pathological variables, including proliferation index, and the model was fit with the tree-structured survival analysis (TSSA) as described previously.¹⁰ This variant is based in the same recursive partitioning algorithm adapted to censored data and uses the log-rank test statistic for splitting. To assess the amount of overfitting, 1,000 10-fold cross-validation experiments were performed as suggested.¹¹ In each of those 1,000 experiments, the data set was randomly split into smaller data sets and a pruning method was used to choose the best number of nodes for the tree. A restriction was imposed on the tree construction such that terminal subgroups resulting from any given split must have at least five patients.

Patient survival was measured from the time of surgery, and the survival distribution for each terminal group was estimated using the product limit method of Kaplan–Meier. Multivariate analysis of survival was also performed using Cox proportional hazards regression analysis. Statistical significance for the differences in survival was tested by the log-rank test based on a 500 permutation-based bootstrap, and defined by a resulting *P* value <0.05.

The sensitivity and specificity of the resulting tree analysis were evaluated using the above-mentioned 10-fold cross validation. CART results for the prediction of death were also compared with the results derived from a model based on logistic regression using the area under the curve (AUC) from the receiver operator characteristic curves. Analyses were performed using S-PLUS software (Version 6.0, Insightful Corp., Seattle, WA, USA).

RESULTS

The demographic and key tumor-related characteristics of the 106 patients analyzed are outlined in

Table 1. Metastases to lymph nodes were most frequent, followed by liver, and lung metastases. Pathological subclassification revealed that all tumors were adenocarcinomas. Histological subclassification based on light microscopy revealed that poorly differentiated tumors were diagnosed in 2.8% of the cases, 42 (39.6%) were moderately differentiated and 61 (57.6%) were well-differentiated tumors.

Proliferation index ranged from 14 to 82%, with a mean (SD) value of 38% (15.4). The median overall survival for the cohort was 5.6 years (95% CI: 5–7.5 years), with 73% surviving at 5 years.

CART Survival Analysis

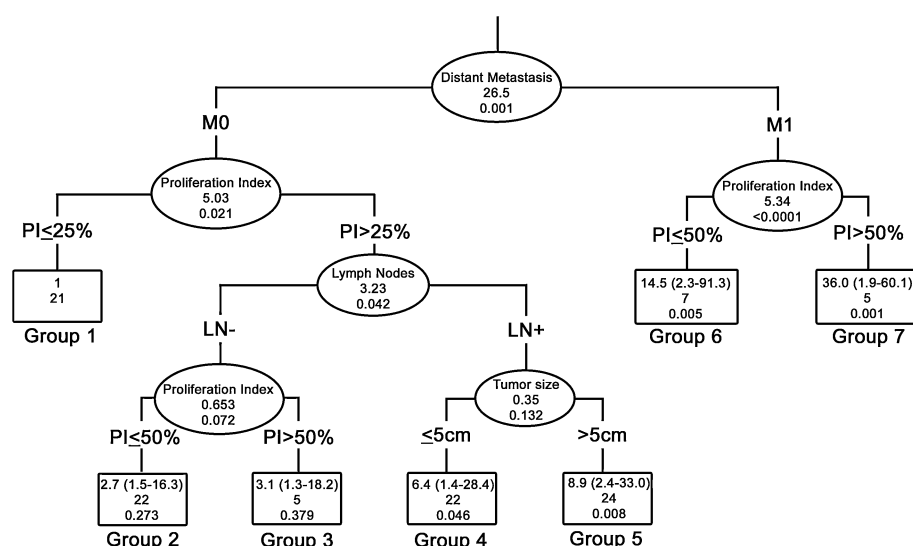
The default tree was generated by allowing the program to determine the variable with the optimal first split. The results from the best tree indicated that distant metastatic involvement was chosen as the initial split 76% of the times, PI was selected with a 27% probability, and lymph node involvement was selected with a 23% probability. The next highest probability was 3%. The default tree, therefore, had an initial split on distant metastasis, and seven terminal nodes or subgroups were formed, with an overall misclassification rate of 16% (17 of 106). The complete model was found to have an overall sensitivity of 84% and a specificity of 82%.

The variables determining the structure of the tree included distant metastasis, PI, lymph node involvement and tumor size, with PI being chosen more than once for splitting other covariates. The structure of the default tree is presented in Fig. 1, and the corresponding survival curves from four of the seven terminal groups generated are presented in Fig. 2. The longest surviving subgroup (group 1) included 21 (19.8%) CRC patients without metastatic organ sites and a PI less than 25%, with no deaths occurring within this group. A second subgroup (group 2) with a relatively long median survival of 5.5 years included 22 patients without distant metastasis, with PI between 25 and 50% and without lymph node involvement. The shortest surviving subgroup (group 7) included CRC patients with liver metastases and a PI higher than 50%. These patients had a median survival of only 3.6 months (95% CI: 0–7.3 months). A subgroup of 24 patients (group 5) without distant metastasis, PI higher than 25%, positive lymph node metastasis and a primary tumor larger than 5 cm also had a short median survival of 36 months.

The results generated by the CART algorithm were then compared with a regression model created using

TABLE 1. Prognostic factors analyzed by the CART algorithm

Variable	Range	Coding	n (%) / [median]	Interpretation
Age	28–90	Continuous	106 [65]	Age at surgery (in years)
Sex	1–2	0: Male 1: Female	60 (56.6) 46 (43.4)	
Location	1–2	0: Colon 1: Rectum	63 (59.4) 43 (40.6)	Tumor location
T category	1–4	0: 1–2 1: 3–4	23 (21.6) 83 (78.4)	Local tumor invasion
Size	1.6–18.5	Continuous	106 [5]	Tumor size (in cm)
Lymph node metastasis	Yes/no	0: No 1: Yes	54 (50.9) 52 (49.1)	Lymph node metastasis
Metastasis	Yes/no	0: No 1: Yes	94 (88.7) 12 (11.3)	Distant metastasis
Differentiated	1–3	0: = Well 1: = Mod 2: = Poor	61 (57.6) 42 (39.6) 3 (2.8)	Tumor differentiation
PI	14–82	Continuous	106 [38]	Ki-67 proliferation index (%)
Lymph	Yes/no	0: No 1: Yes	66 (62.3) 40 (37.7)	Microscopic lymphatic invasion
Venous	Yes/no	0: No 1: Yes	70 (66.0) 36 (34.0)	Microscopic venous invasion

**FIG. 1.** Pruned survival tree. Inside each *ellipse* (intermediate nodes) is the log-rank test statistic and a permutation *P* value for the split. Inside each terminal node (*boxes*) is the hazard ratio (95% CI) relative to the left most node in the tree, the number of cases in the node, and the corresponding *P* value.

the same factors. These analyses were performed using a randomly splitted 70% of the sample as a training and validation set and the other 30% as the testing data set. The CART decision tree had a receiver operator characteristic curve AUC of 0.74. When the same variables were entered into the regression model, the AUC was 0.78 ($P = 0.416$).

The Cox proportional hazards model confirmed the findings of the Ki-67-based PI as an independent factor of prognosis in the dataset (Table 2).

DISCUSSION

Prognostic groups for patients with CRC have been defined primarily based on the extent of the disease as

categorized by specific clinical and pathological characteristics.¹² Some of these patients, however, fall outside of the standard subsets generated by the classifying algorithms and therefore, current clinical research efforts have focused on identifying patient subsets that may be responsive or may need further therapy after surgery for colorectal cancer.¹³ Also because of the complex presentations of patients with CRC, clinicians often experience difficulty applying standard statistical methods to assess the interactions between clinical variables, determining the cumulative effect of these variables on survival, and translating this information into appropriate management.

In this study, a cellular marker of proliferation, the protein Ki-67, was evaluated by immunohisto-

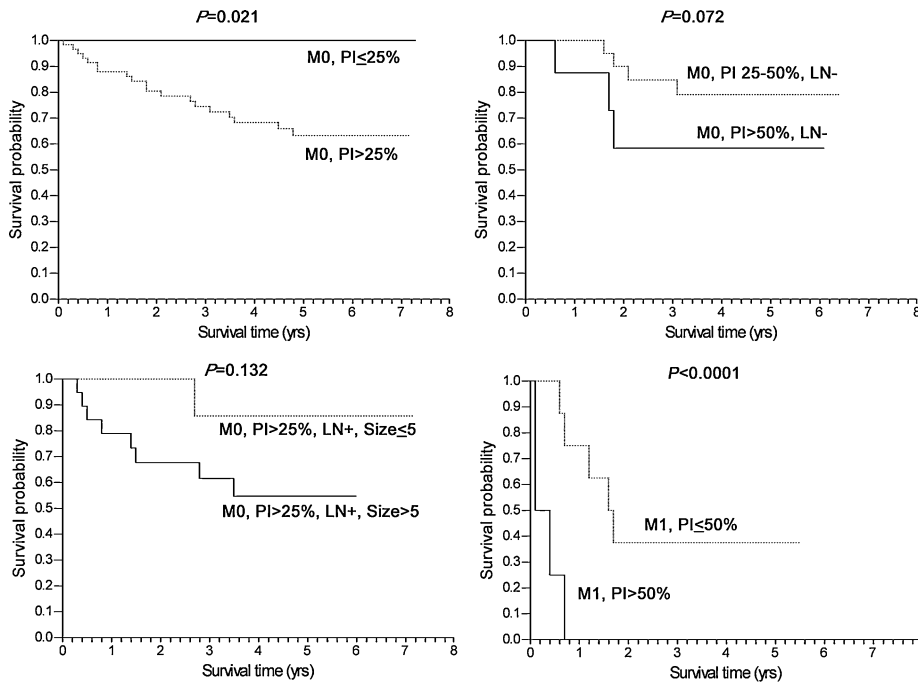


FIG. 2. Survival curves for four of the terminal nodes generated by the CART algorithm. The proliferation index helped to discriminate TNM stages II–IV individuals into different subset of patients and survival probabilities.

TABLE 2. Multivariate analysis of significant factors associated with overall survival in colorectal cancer

Variable	Regression coefficient	SE	Hazard ratio	95% CI	P-value ^a
T category					
T3–T4 versus T1–T2	0.638	0.33	1.9	1.0–3.8	0.039
Lymph node metastasis					
Positive versus negative	1.212	0.52	3.4	1.21–9.3	0.019
Distant metastasis					
Present versus absent	2.195	0.54	8.9	3.1–26.1	<0.001
Proliferation index (continuous variable)	0.735	0.34	2.1	1.1–4.1	0.032

^a Cox regression.

chemistry in CRC samples from patients who underwent surgical treatment. Classification and regression tree (CART) analysis was then used to assess the value of the protein in predicting patient survival and to explore the interaction with other common covariates.

Using a straightforward and relatively inexpensive immunohistochemical technique to evaluate tumor cell proliferation, the results demonstrated not only that the Ki-67 protein is useful for predicting outcome after surgery for colorectal cancer, but also identified previously unappreciated patient subsets inside the more established TNM classification based on the interaction between important common clinicopathologic covariates and this biological marker.

The results on proliferation index are in agreement with earlier observations demonstrating higher proliferative activity in more advanced colorectal

tumors.¹⁴ While a direct comparison of studies is limited, the proliferation indices reported here lie within the 7–94% reported for colorectal carcinomas in the literature.¹⁵

Cell proliferation, as a summarizing feature of an altered cell cycle and underlying cell signaling in cancer, has been shown to be a powerful indicator of tumor behavior without the need of analyzing every factor involved in the pathway.¹⁶ This has been demonstrated in several solid tumors including colorectal adenocarcinomas, where an increased proliferating fraction of tumor cells, as identified by the protein Ki-67, was related with an increase in the risk of death after surgery.⁵

Proliferation index appeared not only as a significant independent factor of prognosis, but also was used more than once by the partitioning algorithm to select groups with well-defined differences in outcome. The main groups identified resulted in

patients with distant metastasis (stage IV) and mid- or low-proliferation indices ($PI < 50\%$) whose long-term survival could be defined as “less poor”, and whom may potentially benefit of more aggressive therapies. Another subgroup was formed by patients without positive lymph nodes (stage II) and tumors with PI over 25% that can be considered to have lower survival probabilities and could possibly benefit from adjuvant therapy in clinical trials settings, since adjuvant therapy for this stage II group is not currently indicated.¹⁷ A third group that is important to mention was formed by patients without distant metastasis, with positive lymph nodes (stage III), PI between 25 and 50%, and small tumors (< 5 cm) with better 5-year survival, that may eventually receive less intensive or modified adjuvant treatment.

CART algorithm has been successfully applied in medicine since its conception in the early 1990s, but its use in predicting outcome in cancer seems to have been subutilized. In oncology, the technique has been useful for classifying patients with unknown primary carcinomas,¹⁸ for predicting long-term survival after surgery for melanoma,¹⁹ for the detection of prostate cancer,²⁰ and for predicting tumor response to preoperative radiotherapy in rectal tumors.²¹ Other applications in medicine include the prediction of hip fractures,²² outcome after acute liver failure,²³ in-hospital mortality in acutely decompensated heart failure,²⁴ and outcome after severe head injury.²⁵

One of the main advantages of the method in comparison with more established survival analysis methods is the ease for the visualization of the interaction between survival covariates, without being less accurate in prediction as demonstrated by the results. Both CART- and logistic regression-based models were comparable in their effectiveness in discriminating good or poor 5-year outcome after surgery.

The use of the tree-structured survival analysis is intended to extend the current application of the standard statistical methods to offer a better understanding on how the complex long-term outcome after surgery for CRC can be modeled and thus changed by intervention. In this sense and based in our results, CART appeared as a useful tool for dissecting complex clinical situations and identifying homogeneous patient populations that can assist clinicians in making a further treatment decisions and prognostic inference in CRC patients for future clinical trials.

ACKNOWLEDGMENTS

Dr. Valera is in receipt of a grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology. The authors thank Mr. Takashi Hatano for his technical assistance.

REFERENCES

1. Sobin LH. TNM: evolution and relation to other prognostic factors. *Semin Surg Oncol* 2003; 21:3–7.
2. Compton CC. Colorectal carcinoma: diagnostic, prognostic, and molecular features. *Mod Pathol* 2003; 16:376–88.
3. Nicholl ID, Dunlop MG. Molecular markers of prognosis in colorectal cancer. *J Natl Cancer Inst* 1999; 91:1267–9.
4. Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer* 2005; 5:845–56.
5. Valera V, Yokoyama N, Walter B, Okamoto H, Suda T, Hatakeyama K. Clinical significance of Ki-67 proliferation index in disease progression and prognosis of patients with resected colorectal carcinoma. *Br J Surg* 2005; 92:1002–7.
6. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med* 2003; 26:172–81.
7. Segal MR, Bloch DA. A comparison of estimated proportional hazards models and regression trees. *Stat Med* 1989; 8:539–50.
8. LeBlanc M, Crowley J. A review of tree-based prognostic models. *Cancer Treat Res* 1995; 75:113–24.
9. LeBlanc M. Tree-based methods for prognostic stratification. In: Crowley J ed. *Handbook of statistics in clinical oncology*. New York: Marcel Dekker (2002) pp 457–72.
10. Gordon L, Olshen RA. Tree-structured survival analysis. *Cancer Treat Rep* 1985; 69:1065–9.
11. Segal MR. Features of tree-structured survival analysis. *Epidemiology* 1997; 8:344–6.
12. Eschrich S, Yang I, Bloom G, et al. Molecular staging for survival prediction of colorectal cancer patients. *J Clin Oncol* 2005; 23:3526–35.
13. Horton JK, Tepper JE. Staging of colorectal cancer: past, present, and future. *Clin Colorectal Cancer* 2005; 4:302–12.
14. Chen YT, Henk MJ, Carney KJ, et al. Prognostic significance of tumor markers in colorectal cancer patients: DNA index, S-phase fraction, p53 expression, and Ki-67 index. *J Gastrointest Surg* 1997; 1:266–73.
15. Daidone MG, Costa A, Silvestrini R. Cell proliferation markers in human solid tumors: assessing their impact in clinical oncology. *Methods Cell Biol* 2001; 64:359–84.
16. Evan GI, Vousden KH. Proliferation, cell cycle and apoptosis in cancer. *Nature* 2001; 411:342–8.
17. Iacopetta B. A biological perspective on the selection of stage II colorectal cancer patients for adjuvant chemotherapy. *Ann Oncol* 2002; 13:1510.
18. Hess KR, Abbruzzese MC, Lenzi R, Raber MN, Abbruzzese JL. Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. *Clin Cancer Res* 1999; 5:3403–10.
19. Averbook BJ, Fu P, Rao JS, Mansour EG. A long-term analysis of 1018 patients with melanoma by classic Cox regression and tree-structured survival analysis at a major referral center: Implications on the future of cancer staging. *Surgery* 2002; 132:589–602; [Discussion 02–4].

20. Garzotto M, Beer TM, Hudson RG, et al. Improved detection of prostate cancer using classification and regression tree analysis. *J Clin Oncol* 2005; 23:4322–9.
21. Zlobec I, Steele R, Nigam N, Compton CC. A predictive model of rectal tumor response to preoperative radiotherapy using classification and regression tree methods. *Clin Cancer Res* 2005; 11:5440–3.
22. Jin H, Lu Y, Harris ST, et al. Classification algorithms for hip fracture prediction based on recursive partitioning methods. *Med Decis Making* 2004; 24:386–98.
23. Baquerizo A, Anselmo D, Shackleton C, et al. Phosphorus as an early predictive factor in patients with acute liver failure. *Transplantation* 2003; 75:2007–14.
24. Fonarow GC, Adams KF Jr, Abraham WT, Yancy CW, Boscardin WJ. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *JAMA* 2005; 293:572–80.
25. Rovlias A, Kotsou S. Classification and regression tree for prediction of outcome after severe head injury using simple clinical and laboratory variables. *J Neurotrauma* 2004; 21:886–93.