Classification and Regression Tree (CART) Analysis in Biomedical Research

Jason S. Haukoos, MD, MS
Department of Emergency Medicine
Denver Health Medical Center
Denver, Colorado
March 23, 2004
Introduction

- A significant proportion of biomedical research is observational.
- Multivariable analyses have become increasingly important in clinical research.
- Parametric regression techniques have historically been the root for analyzing multivariable problems.
Introduction

• Specific research questions lend themselves to nontraditional multivariable techniques.
• CART (a non-parametric multivariable technique), in these settings, is used to classify patients into clinically-important categories, especially when outcomes are rare.
Introduction

Demonstrate the use of CART in three distinct clinical settings.

1. Emergency department triage of patients infected with HIV.
2. Survival prediction of patients with colon and rectal cancer.
Emergency Department Triage of Patients Infected with HIV

Jason S. Haukoos, MD, Mallory D. Witt, MD, Christel M. Zeum, MD,
Thomas J. Lee, MD, MSHS, John D. Halamka, MD, MS,
Roger J. Lewis, MD, PhD

Abstract
Objective: The emergency department (ED) and HIV specialty clinics are primary sources of care for persons infected with HIV. HIV disease may be complicated by vague and complex symptomatology, and determining the degree of illness at triage is often difficult. The goals of this project were to characterize the ED presentation of HIV-related conditions, to develop a clinical decision rule to triage HIV-infected patients, and to validate the rule in clinical practice. Methods: The study population consisted of ambulatory patients with self-reported HIV infection who presented for care to the ED of a 553-bed public hospital that serves a medically indigent, minority population. An Illness Severity Instrument was developed by an expert panel to serve as the criterion standard for defining medical urgency for HIV-infected patients presenting to the ED for care. Two phases of the study were conducted: data from the first phase, a non-interventional cohort study, were used to develop a clinical decision rule for the ED triage of HIV-infected patients. The second phase was a prospective validation of the clinical decision rule. Results: During phase I, data from 542 patient visits were collected. Data from 441 (81%) patient visits were used in a classification and regression tree (CART) analysis to produce a decision rule, the Clinical Triage Instrument. During phase II, the prospective validation of the Clinical Triage Instrument, 156 patient visits occurred. Of these, 86 (56%) patient visits were triaged using the Clinical Triage Instrument and could be scored using the Illness Severity Instrument. The Clinical Triage Instrument accurately triaged 45 (51%; 95% confidence interval (95% CI) = 40% to 62%) patient visits, undertriaged 11 (13%; 95% CI = 6% to 21%) patient visits, and overtriaged 32 (36%; 95% CI = 26% to 47%) patient visits. Sensitivities and specificities for determining emergent, urgent, and nonurgent medical conditions by the Clinical Triage Instrument were 56% (95% CI = 31% to 75%) and 84% (95% CI = 64% to 92%), 71% (95% CI = 53% to 84%) and 39% (95% CI = 25% to 55%), and 18% (95% CI = 6% to 32%) and 93% (95% CI = 84% to 98%), respectively. The positive and negative predictive values for determining an emergent medical condition using the Clinical Triage Instrument were 48% (95% CI = 26% to 70%) and 88% (95% CI = 78% to 95%), respectively. The positive and negative predictive values for determining a nonurgent medical condition using the Clinical Triage Instrument were 56% (95% CI = 21% to 86%) and 71% (95% CI = 50% to 81%), respectively. Conclusions: The Clinical Triage Instrument was not sufficiently accurate for clinical use. Until accurate and reliable triage methods are developed, all patients infected with HIV who present to the ED for care should receive timely evaluation and care. Key words: HIV; triage; clinical decision rule; Illness Severity Instrument; Clinical Triage Instrument. ACADEMIC EMERGENCY MEDICINE 2002; 9:880–888.
HIV Triage

• Background
  – HIV-infected patients frequently seek care in emergency departments (EDs).
  – HIV is complicated by vague and complex symptomatology.
  – Determining the degree of illness, and thus the level of care needed, at triage is difficult.
HIV Triage

• Objectives
  – To characterize the ED presentation of HIV-related conditions.
  – To develop a clinical decision rule to triage HIV-infected patients.
  – To validate the rule in clinical practice.
HIV Triage

• Methods
  – Study population consisted of ambulatory patients with self-reported HIV infection who presented to the ED at Harbor-UCLA Medical Center in Torrance, California.
  – Harbor-UCLA Medical Center is a 553-bed public teaching hospital in Los Angeles County with an annual ED census of ~ 70,000.
HIV Triage

• Methods
  – The first phase consisted of collecting data from approximately 500 consecutive HIV-infected patients.
  – These data were used to develop the clinical decision rule for the ED triage of HIV-infected patients.
HIV Triage

• Methods
  – Potential Predictor Variables: chief complaint, CD4 count (if known), symptoms (including fever, chills, sweats, worsening headache, confusion, ataxia, change in vision, dizziness upon standing, dysphagia, vomiting, diarrhea, cough, hemoptysis, dyspnea at rest or with exertion, and new rash), and vital signs.
HIV Triage

• Methods
  – **Outcomes Variables:** Emergent, Urgent, or Non-urgent using a validated Illness Severity Instrument (ISI).
  – The ISI used vital signs, results of the physical examination, laboratory evaluation, ED diagnosis, and disposition to classify patients into one of the three categories.
HIV Triage

• Methods
  – Part I of this study consisted of developing a clinical triage instrument (CTI) that would divide patients into risk groups based upon information obtained at triage.
  – Part II of this study consisted of prospectively validating the derived rule on an addition set of patients.
HIV Triage

• Results of Part I
  – Of 542 HIV-infected patients who presented to the ED for care, 441 (81%) had documentation that allowed scoring using the ISI.
  – Preliminary CART analysis (not shown) created a decision rule that had a significant rate of mis-triage using cross-validation.
HIV Triage

• Results of Part I
  – Most of those under-triaged were patients with altered mental status (AMS), bleeding, traumatic injuries, or non-HIV-related disease.
  – Those with AMS, bleeding, traumatic injuries and non-HIV-related disease were excluded and CART analysis was performed on the remaining 390 patients.
<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Target</th>
<th>Predictor</th>
<th>Weight</th>
<th>Categorical</th>
<th>Aux.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SX_SOB_AT_REST$</td>
<td></td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>SX_SOB_EXERCISE$</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>SX_CP$</td>
<td></td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>SX_RASH$</td>
<td></td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>TRIAGE_CAT$</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>TEMP</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>PULSE</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>RESP</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>SBP</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
</tbody>
</table>

Sort: File Order
Select Predictors
Select Categorical
Select Aux.

Tree Type
- Classification
- Regression

Set Focus Class...
Target Variable
VISIT_CAT$
Weight Variable

Number of Predictors: 33

Save Group... CART Combine Score... Cancel Continue Start
Select Method for Testing Tree

- No independent testing - exploratory tree
- V-fold cross-validation: 100
- Fraction of cases selected at random for testing: 0.200000
- Test sample contained in a separate file: 
- Variable separates learn and test samples:
  STUDY_PT_NUM
  HH_NO$
  RIPT$
  DATE
  COMMENT$
  CD4_DATE
### Costs for VISIT_CAT$

<table>
<thead>
<tr>
<th>Level</th>
<th>EM</th>
<th>NU</th>
<th>UR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>NU</td>
<td>6</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>UR</td>
<td>1.000</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

**Costs Per Level**

<table>
<thead>
<tr>
<th>Level</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>NU</td>
<td></td>
</tr>
<tr>
<td>NU</td>
<td></td>
</tr>
</tbody>
</table>
### Priors For: VISIT_CAT$

- **EQUAL**: All categories have equal probability
- **LEARN**: Probabilities match learn sample frequencies
- **TEST**: Probabilities match test sample frequencies
- **DATA**: Probabilities match total sample frequencies
- **MIX**: The average of EQUAL and DATA
- **SPECIFY**: Specify priors for each level

#### Specify Priors

<table>
<thead>
<tr>
<th>Class</th>
<th>Priors</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>1</td>
</tr>
<tr>
<td>NU</td>
<td>1</td>
</tr>
<tr>
<td>UR</td>
<td>1</td>
</tr>
</tbody>
</table>
HIV Triage

• Results of Part II
  – 88 patients were studied.
  – The positive predictive value for determining an emergent medical condition using the CTI was 48% (95% CI: 26% - 70%).
  – The negative predictive value for determining a non-urgent medical condition using the CTI was 71% (95% CI: 60% - 81%).
**HIV Triage**

- Conclusions
  - Triage of HIV-infected patients is difficult.
  - Attempts to validate triage methods are complicated by the absence of a definition of medical urgency for this population.
  - The CTI was not sufficiently accurate to be used when triaging patients infected with HIV.
Carcinoembryonic Antigen and Albumin Predict Survival in Patients With Advanced Colon and Rectal Cancer

Matthew R. Dixon, MD; Jason S. Haukoos, MD, MS; Sejal M. Udani; Jesse J. Naghi; Tracey D. Arnell, MD; Ravin R. Kumar, MD; Michael J. Stamos, MD

**Hypothesis:** Patients with stage IV colon or rectal cancer at initial diagnosis have characteristics that will predict subsequent survival time.

**Design:** Retrospective cohort study.

**Setting:** Urban county teaching hospital providing tertiary care.

**Patients:** Patients who came to the study institution with stage IV colon or rectal cancer between 1991-1999.

**Main Outcome Measure:** Survival duration (days) after diagnosis.

**Results:** One hundred five patients were identified, with a median survival of 225 days (interquartile range, 72-688 days). Univariate analysis identified carcinoembryonic antigen (CEA) and albumin (ALB) as possible predictors for survival. Classification and regression tree analysis, a form of binary recursive partitioning, was used to identify optimal cut points for CEA (275 ng/mL) and ALB (2.7 g/dL) levels. Based on the cut points, patients were stratified into the following groups: (1) low CEA, high ALB; (2) low CEA, low ALB; (3) high CEA, high ALB; and (4) high CEA, low ALB. The median survival times for the first group and the fourth group were 287 days (interquartile range, 150-851 days) and 39 days (interquartile range, 14-168 days), respectively. A Kaplan-Meier analysis was performed, and a statistically significant difference was identified across all strata (P = .004). Additionally, groups 1 and 4 demonstrated the largest overall survival difference (P < .001).

**Conclusions:** Patients with stage IV colon and rectal cancer with a CEA level greater than or equal to 275 ng/mL and an ALB level less than 2.7 g/dL had a significantly shorter survival time. Conversely, patients with an ALB level greater than or equal to 2.7 g/dL and a CEA level less than 275 ng/mL had a longer survival time.

Arch Surg. 2003;138:962-966
Cancer Survival Prediction

• Background
  – Colorectal cancer is the second leading cause of cancer death in the United States.
  – Approximately 130,000 patients are diagnosed with this form of cancer annually.
  – When diagnosed, approximately 25% will have evidence of metastatic disease.
Cancer Survival Prediction

- Background
  - Patients with stage IV colorectal cancer display survival heterogeneity.
  - Identifying patients with short and long survival times may significantly impact the process and type of care provided.
Cancer Survival Study

- Objectives
  - To identify characteristics of patients with stage IV colorectal cancer at the time of diagnosis that would help separate patients into groups with distinct survival probabilities.
Cancer Survival Study

• Methods
  – Retrospective cohort study performed at Harbor-UCLA Medical Center in Torrance, California.
  – Consecutive patients identified by the tumor registry with stage IV colorectal cancer were enrolled between 1991 – 1999.
Cancer Survival Study

• Methods
  – Potential Predictor Variables: age, sex, race/ethnicity, initial symptoms (weight loss, obstruction, rectal bleeding, pain, and constipation), initial laboratory values (hematocrit, MCV, creatinine, PT, AST, ALT, albumin, total bilirubin, CEA, alkaline phosphatase, and fibrinogen), and location of the primary tumor and metastases.
Cancer Survival Study

• Methods
  – Other Variables: whether surgery, radiation, or chemotherapy was performed.
  – Outcome Variables: length of survival following diagnosis (in days).
Cancer Survival Study

• Results

– 105 patients were included.
– 99 patients had survival follow-up data with a median survival time of 225 (IQR: 72 – 688) days.
– Patients were categorized into “long” and “short” survival groups with a cutoff point of 120 days.
Cancer Survival Prediction

• Results
  – Using Wilcoxon rank sum tests, albumin and CEA were found to be statistically significant with respect to the categorized survival variable.
  – CART was then used to find optimal cutoff points for albumin and CEA in order to split the dataset most homogeneously with respect to the outcome variable.
<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Target</th>
<th>Predictor</th>
<th>Weight</th>
<th>Categorical</th>
<th>Aux.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTT</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INR</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALBUMIN</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIBRINOGEN</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEA</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CXR$</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT_DATE$</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRI$</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BDONESCA$</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tree Type: ✔️ Classification  
Weight Variable: SURVIVAL_CATEGORY$  
Number of Predictors: 2
### Model Setup

**Select Method for Testing Tree**

- No independent testing - exploratory tree
- **V-fold cross-validation:** 100
- Fraction of cases selected at random for testing: 0.200000
- Test sample contained in a separate file: 
- Variable separates learn and test samples:

  - PT_  
  - PT_ID_  
  - SEX$  
  - RACE$  
  - DOB  
  - DATE_OF_DIAG

Save Group...  CART  Combine  Score...  Cancel  Continue  Start
### Costs for SURVIVAL_CATEGORY$

<table>
<thead>
<tr>
<th>Level</th>
<th>Long</th>
<th>Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long</td>
<td></td>
<td>2.000</td>
</tr>
<tr>
<td>Short</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Cost: 2.000
### Priors for: SURVIVALCATEGORY$

- **EQUA**: All categories have equal probability
- **LEARN**: Probabilities match learn sample frequencies
- **TEST**: Probabilities match test sample frequencies
- **DATA**: Probabilities match total sample frequencies
- **MIX**: The average of EQUA and DATA
- **SPECIFY**: Specify priors for each level

#### Specify Priors

<table>
<thead>
<tr>
<th>Class</th>
<th>Priors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long</td>
<td>1</td>
</tr>
<tr>
<td>Short</td>
<td>1</td>
</tr>
</tbody>
</table>
Cancer Survival Prediction

• Results
  – Using the cutoff points for albumin and CEA, we categorized patients into one of four groups.
    • Group 1 – High albumin/Low CEA
    • Group 2 – Low albumin/Low CEA
    • Group 3 – High albumin/High CEA
    • Group 4 – Low albumin/High CEA
Conclusions

- A significant survival difference was demonstrated among the four groups.
- These results may help the clinician in determining the appropriate counseling and optimal treatment for patients with advanced colon and rectal cancer.
Prediction Rules for Estimating Neurologic Outcome Following Out-of-Hospital Cardiac Arrest

Jason S. Haukoos, M.D., M.S.¹-⁵
Roger J. Lewis, M.D., Ph.D.³-⁵
James T. Niemann, M.D.³-⁵

Author affiliations are listed on the following page.


Supported, in part, by an Individual National Research Service Award from the Agency for Healthcare Research and Quality (F32 HS11509) and a Research Training Grant from the Society for Academic Emergency Medicine to Dr. Haukoos.
Cardiac Arrest Survival

• Background
  – Approximately 300,000 out-of-hospital cardiac arrests occur annually in the United States.
  – Many factors are thought to be related to survival including activating EMS, performing CPR and defibrillation, the initial rhythm, whether the arrest was witnessed, its location, the patient’s age, and his or her comorbidities.
Cardiac Arrest Survival

• Background
  – Survival to hospital discharge (SHD) ranges from 1% to 40% depending upon the population studied.
  – Unfortunately, a significant proportion of survivors have irreversible anoxic brain injury.
  – No model has been developed to predict survival with meaningful neurologic function using variables available to emergency care personnel.
Cardiac Arrest Survival

• Objectives
  – The purpose of this study was to develop clinical decision rules that could accurately predict meaningful survival, defined by the Glasgow Coma Score (GCS) at the time of hospital discharge, following out-of-hospital cardiac arrest.
Cardiac Arrest Survival

• Methods
  – This was a retrospective cohort study of consecutive adult patients treated for out-of-hospital nontraumatic cardiac arrest and transported to Harbor-UCLA Medical Center between 1994 and 2001.
Cardiac Arrest Survival

• Methods
  – **Possible Predictor Variables**: age, sex, race/ethnicity, site of arrest, whether it was witnessed, whether bystander CPR was performed, the initial arrest rhythm, whether an AED was used, patient downtime, and paramedic response time.
  – **Outcome Variables**: SHD with a GCS $\geq 13$, $\geq 14$, and $=15$. 
Cardiac Arrest Survival

• Results
  – 754 total patients were studied.
  – 36 (5%) SHD.
  – 16 (2%) SHD with a GCS ≥ 13.
  – 15 (2%) SHD with a GCS ≥ 14.
  – 5 (0.7%) SHD with a GCS = 15.
Model Setup

Select Method for Testing Tree

- No independent testing - exploratory tree
- V-fold cross-validation: 100
- Fraction of cases selected at random for testing: 0.200000
- Test sample contained in a separate file:
- Variable separates learn and test samples:

```
DATE
HHNUMBER
RACE$
PMH1$
PMH2$
PMH3$
```
Model Setup

Priors For: MSHD13

- EQUAL: All categories have equal probability
- LEARN: Probabilities match learn sample frequencies
- TEST: Probabilities match test sample frequencies
- DATA: Probabilities match total sample frequencies
- MIX: The average of EQUAL and DATA
- SPECIFY: Specify priors for each level

Specify Priors

<table>
<thead>
<tr>
<th>Class</th>
<th>Priors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Actual Class</td>
<td>Total Cases</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>0</td>
<td>738</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
</tr>
</tbody>
</table>
### Test Sample Prediction Success Table

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Total Cases</th>
<th>Percent Correct</th>
<th>0 N=517</th>
<th>1 N=237</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>739</td>
<td>63.95%</td>
<td>517</td>
<td>222</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>100.00%</td>
<td>0</td>
<td>15</td>
</tr>
</tbody>
</table>

Learn Test Class: None

Count Flow Column %
### Test Sample Prediction Success Table

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Total Cases</th>
<th>Percent Correct</th>
<th>N=506</th>
<th>N=246</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>749</td>
<td>67.00</td>
<td>507</td>
<td>242</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>80.00</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Notes:
- The table shows the prediction success for two classes, 0 and 1.
- The percent correct column indicates the percentage of cases that were correctly predicted.
- The table compares the prediction success between two sample sizes, N=506 and N=246.
Cardiac Arrest Survival

• Conclusions
  – This study reported preliminary decision rules for meaningful survival to hospital discharge following out-of-hospital cardiac arrest with high NPVs for each.
  – Future studies need to be performed to prospectively validate these models.
Conclusions

- Three different approaches to using CART for biomedical research.
- CART is a powerful technique with significant potential and clinical utility.
- Use of CART is likely to increase in the future, largely because of the substantial number of important problems for which it is the best available solution.