College of Pharmacy
UIC University of Illinois at Chicago

Comparing Exhaustive CHAID Classification Tree and Forward Stepwise Logistic Regression in Explaining the Prescribing of Antidepressants

HiangKiat (Jason) Tan[1], M.S., Swu-Jane Lin[2], Ph.D., Bruce L. Lambert[2], Ph.D., Stanley L. Slove[3], Ph.D.

[1]HealthCore, Wilmington, DE, USA, [2]Department of Pharmacy Administration, University of Illinois at Chicago, IL, USA, [3]Information & Decision Science Department, University of Illinois at Chicago, IL, USA

For more information, contact:
HiangKiat (Jason) Tan, M.S.
HealthCore
Health Outcomes Research
800 Delaware Avenue, Fifth Floor,
Wilmington, DE, USA
Tel: 302-230-2132
E-mail: jtan@healthcore.com

BACKGROUND

- The popularity of classification tree has led to an increase in the number of empirical comparisons between classification trees and other statistical classifiers, such as logistic regression, on a variety of problems.
- However, the performance of classification trees and logistic regression varied across datasets.
- The performance of both approaches strongly depends on some general features of the datasets, such as number of covariates, type of covariates, and distribution of the variables. No single rule has been able to guide the choice between the methods.
- While Classification tree has been widely used in other disciplines, relatively few is known about its performance in medical field particularly in explaining prescribing behavior.
- More comparisons are needed to better understand their performance in different contexts.

OBJECTIVE

- To compare and contrast the Exhaustive CHAID (Chi-square Automatic Interaction Detection) classification tree, a algorithm commonly used in marketing research, with forward stepwise logistic regression (LR) in explaining the prescribing of antidepressants.
  - Performance was evaluated by the identified explanatory variables and interaction effects, correlation of estimated probabilities, classification accuracy, sensitivity, specificity, and curves of Receiver Operating Characteristic (ROC).

METHODS

- Data: 1997~2001 National Ambulatory Medical Care Survey (NAMCS)
- Inclusion criteria: Office visits with complete data.
- Dependent variables: the prescribing of antidepressants (Yes/No)
- Explanatory variables:
  - Binary variable: patient gender, whether the physician had seen the patient before (old/new patient), whether the physician was the patient's primary care physician (PCP), whether the patient reported any depressive symptoms, whether the patient was diagnosed with depression, whether the physician practiced independently or in collaboration (solo/non-solo), and the location of the physician's practice (MSA/non-MSA).
  - Categorical variable (more than 2 categories): physician's specialty, patient's race and payment source, and census region of a physician's practice (e.g. Northeast, etc.)
  - Continuous variable: patient's age, duration of a visit.
- The data was randomly divided into a training set and a test set with a 7 to 3 ratio.
- Training set was used to train the Exhaustive CHAID and forward stepwise logistic regression models.
- The test set was used to evaluate the performance of both models in terms of correlation of logits, classification accuracy, sensitivity, specificity, and ROC curve.

RESULTS

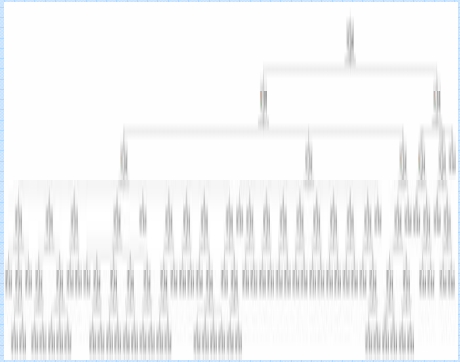
- 113,128 office visits met the inclusion criteria.
- About 6.7% of the visits were prescribed with at least one antidepressant between 1997 and 2001.
- There were 79,294 (70.09%) in the training set and 33,834 (29.01%) in the test set.

Forward Stepwise LR Model

Variables	Odds Ratio	95% Confidence interval
Age (year)	1.013**	(1.011, 1.015)
Gender: Male (ref)	-	-
Female	1.572**	(1.461, 1.691)
Race: White (ref)	-	-
Black	0.763**	(0.667, 0.874)
Others	0.668**	(0.538, 0.829)
PCP: Yes	1.359**	(1.203, 1.535)
Payment: Private (ref)	-	-
Medicare	0.762**	(0.687, 0.844)
Medicaid	Did not enter	
Self-pay	0.557**	(0.487, 0.637)
Others	0.784**	(0.681, 0.902)
Seen before: Yes	1.236**	(1.099, 1.391)
MSA: Yes	0.905*	(0.824, 0.994)
Region: South (ref)	-	-
Northeast	0.896*	(0.821, 0.978)
Midwest	0.850**	(0.776, 0.931)
West	Did not enter	
Solo practice: Yes	0.862**	(0.800, 0.928)
Specialty: PCP (ref)	-	-
Psychiatry	12.591**	(10.749, 14.748)
Others	0.737**	(0.651, 0.834)
Reported symptom: Yes	1.653**	(1.438, 1.900)
Being diagnosed: Yes	9.966**	(8.880, 11.185)
Duration of visit	1.006**	(1.004, 1.009)

*Significant at p<0.05 level.
**Significant at p<0.001 level.

Exhaustive CHAID Model



- While the forward stepwise LR resulted to all 13 explanatory variables as significant, the Exhaustive CHAID identified 11 explanatory variables and 3 interactions as significantly associated with the prescribing of antidepressants.

Terminal nodes in Exhaustive CHAID Tree

Terminal Nodes Description	Node ID	Number of Visits	Prescribing Rate (%)
Diagnosed, private insurance, psychiatrist, time<=30	72	564	85.8
Diagnosed, Medicare/Medicaid/Others, psychiatrist, no symptom	74	277	76.9
Not diagnosed, other specialties, 10<time<=14, non-solo	40	614	0.7
Not diagnosed, other specialties, time<=9, age<=40	34	1914	0.5
Not diagnosed, other specialties, 14<time<=15, age<=8	41	495	0.4

- In Exhaustive CHAID model, each subgroup had different set of explanatory variables associated with the dependent variable. For example, node 72 (all 564 patients who were diagnosed with depression, having private insurance, and visit psychiatrist within 30 minutes) used four explanatory variables to explain and predict the prescribing of antidepressants.
- On the other hand, LR uses all significant explanatory variables to associated with the dependent variable for each observation.
- The Pearson correlation of the predicted logits from both models was 0.7649 (p<0.001).

Comparisons between Two Models

Indicator\Model	Forward Stepwise LR (%)	Exhaustive CHAID (%)
Accuracy	0.94887	0.95162
Sensitivity	0.35546	0.50512
Specificity	0.99104	0.98335
Area under ROC curve (AUC)	0.8507 (SD=0.0052)	0.8610 (SD=0.0047)

- At a conventional cutpoint of 0.5, the classification accuracy and specificity were very similar for both models, but the sensitivity of the Exhaustive CHAID was slightly better than that of forward stepwise LR.
- In ROC analysis, the Exhaustive CHAID significantly outperforms forward stepwise LR in AUC comparison (p<0.001).

CONCLUSIONS

- The performance of Exhaustive CHAID was at least as comparable with forward stepwise logistic regression.
- The resulting predicted logits of both models were highly correlated.
- In addition, Exhaustive CHAID has the capacity to automatically detect interaction effects without having to specify *a priori* the potential interaction terms.
- The Exhaustive CHAID produces a more parsimonious model by using fewer variables to explain the dependent variable.
- The resulting classification tree also provides a visually informative structure on how variables are selected into the model by their relative contributions.

Note: This study was based upon a thesis in partial fulfillment of the requirements for the Master degree at the Graduate College of the University of Illinois at Chicago.