Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010), 50-54

Copyright 2010 by Optimal Data Analysis, LLC

2155-0182/10/$3.00

Manual *vs*. Automated CTA: Optimal Preadmission Staging for Inpatient Mortality from *Pneumocystis cariini* Pneumonia

Paul R. Yarnold, Ph.D. and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

Two severity-of-illness models used for staging risk of in-hospital mortality from AIDS-related *Pneumocystis cariini* pneumonia (PCP) were developed using hierarchically optimal classification tree analysis (CTA), with models derived manually via UniODA software. The first of the "*Manual vs. Automated CTA*" series, this study contrasts classification results between original models and corresponding new models derived using automated analysis. Findings provide superior staging systems which may be employed to improve results of applied research in this area.

Software designed to conduct automated CTA became commercially available in the summer of 2010.[1] Research conducted before this time obtained CTA models by a laborious manual process involving UniODA software.[2,3] Beyond obvious savings in time and labor, two primary advantages of automated CTA involve pruning.

First, Type I error for the CTA model is ensured at an investigator-specified level via a sequential Bonferroni procedure.[3] When the CTA model is derived manually, the Bonferroni procedure is conducted as best as possible as the model is grown (this becomes increasingly difficult as the model gains in complexity), as well as after the model can no longer be expanded. Attributes in close proximity to the root variable and having *p* near 0.05, may be forced out of the model as an increasing number of attributes load on lower branches, disrupting the model and the modeling process. When conducting automated analysis however, this recursive trimming and re-development process is user-transparent: the computer simply executes the algorithm.

Second, the automated software always conducts optimal pruning to explicitly maximize model accuracy, another process which becomes difficult to accomplish manually for complex models.[4] This paper illustrates these advantages using data previously assessed by manual CTA.

PCP in the Early AIDS Era

Research with a sample of 1,339 patients hospitalized with HIV-associated PCP between 1987 and 1990—when hospital mortality rates

Copyright 2010 by Optimal Data Analysis, LLC

2155-0182/10/$3.00

ranged as high as 60%, is considered first.[5] With five attributes (alveolar-arterial oxygen gradient, $AaPo_2$; age—used twice; body mass index; and a binary indicator of whether a patient had prior history of AIDS) the manually-derived CTA model correctly classified 34.1% of 205 patients who died, and 87.0% of 988 living patients (146 patients had missing data on some attributes in the model), yielding a relatively weak[2] ESS= 21.2. This model offered an order-of-magnitude gain in ESS versus the best prior linear model (logistic regression), and more than doubled the ESS achieved by the best prior classification tree model (regression-based recursive partitioning).[5] This CTA model was pruned to maximize ESS, correctly classifying 74.6% of dead and 59.1% of living patients, and returning moderate ESS=33.7: a 59% improvement versus the non-optimized model.[4] Using three attributes, efficiency=11.2 ESS units-per-attribute, and thus the optimized model was 165% more efficient than the original model (4.2 ESS units-per-attribute).

Automatic CTA software was used to obtain an enumerated CTA model using the same attributes and data available for prior logistic regression and recursive partitioning analyses (see Figure 1). The enumerated CTA model had 69.5% sensitivity, 70.1% specificity, moderate ESS=39.7 (17.8% greater than for the optimized manual model), and efficiency=13.2 ESS units-per-attribute (17.9% greater than the optimized manual model). Analysis was completed in 278 CPU seconds using a 3 GHz Intel Pentium D microcomputer (used in all analyses).

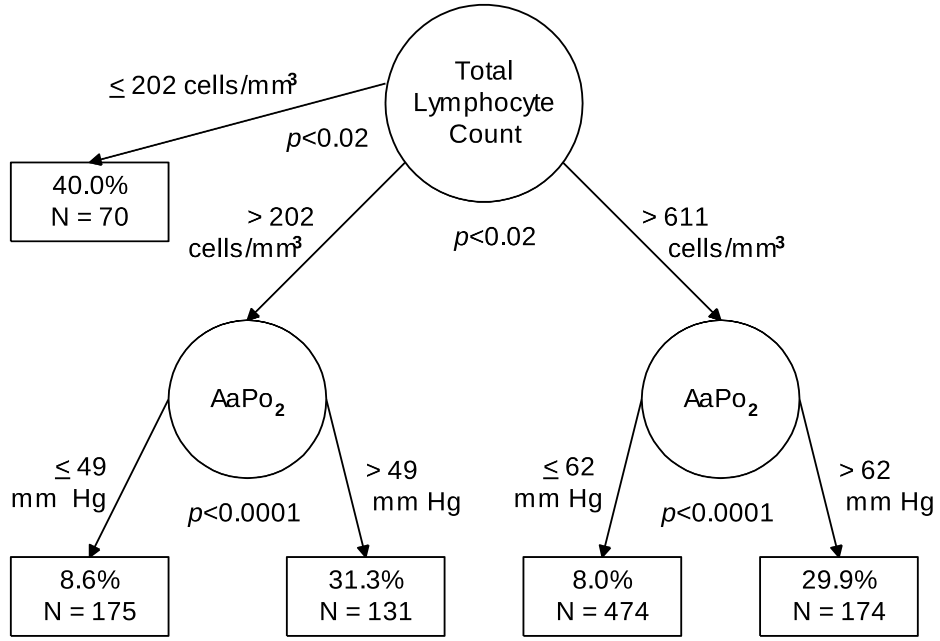

Figure 1: Enumerated CTA Model for Predicting PCP Inpatient Mortality Prior to 1995

Research in the Highly Active Antiretroviral Therapy (HAART) Era

Research investigating a sample of 1,660 patients hospitalized with HIV-associated PCP between 1995 and 1997—the period marking early adoption of non-nucleoside reverse transcriptase and protease inhibitors as HIV therapy, is considered next.[6] Using four attributes (wasting, $AaPo_2$—used twice, and Albumin, the manually-constructed CTA model correctly classified 59.4% of 128 patients who died, and

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

73.7% of 1,066 patients who lived (466 patients had missing data for model attributes), yielding moderate ESS=33.1. Pruned to maximize ESS, the two-attribute optimized model had 53.8% sensitivity (correct prediction of dead patients), 84.3% specificity (correct prediction of living patients), and moderate ESS=45.2 (the optimized model trimmed two nodes previously emanating from the right side of the root node). The optimized model thus offers a 36.6% increase in ESS versus the original model, as well as 172% greater efficiency (22.6 vs. 8.3 ESS units-per-attribute, respectively).[2,4]

An enumerated CTA model was conducted via ODA automatic CTA software, allowing a jackknife-unstable attribute to enter the model if it met the Bonferroni criterion[2] for statistical significance, and if its jackknife ESS exceeded training or jackknife ESS afforded by alternative attributes. To facilitate a direct comparison of models, the three-attribute enumerated model was developed *using the attributes selected by the manually derived model*: wasting, $AaPo_2$, and Albumin. The enumerated CTA model (see Figure 2) had 65.4% sensitivity, 88.2% specificity, a *relatively strong* ESS=53.7 (19% greater than for the optimized manual CTA model), and efficiency=17.9 ESS units-per-attribute (20.8% lower than for the optimized manual model). Analysis was completed in 101 CPU seconds.

In such "disease-staging research" it is customary to provide a *staging table*, such as in Table 1.[5] Rows in the staging table are CTA model endpoints which have been reorganized in order of increasing percent of class 1 (dead patients) membership. Stage is an *ordinal index* of severity of illness, and p_{death} is a *continuous index*: increasing values on either index indicate increasing (worsening) disease severity. The 1st and 4th strata reflect a 16-fold difference in likelihood of dying in-hospital: compared to Stage 1, p_{death} is about four times as high in Stage 2, fifteen times as high in Stage 3, and sixteen times as high in Stage 4.

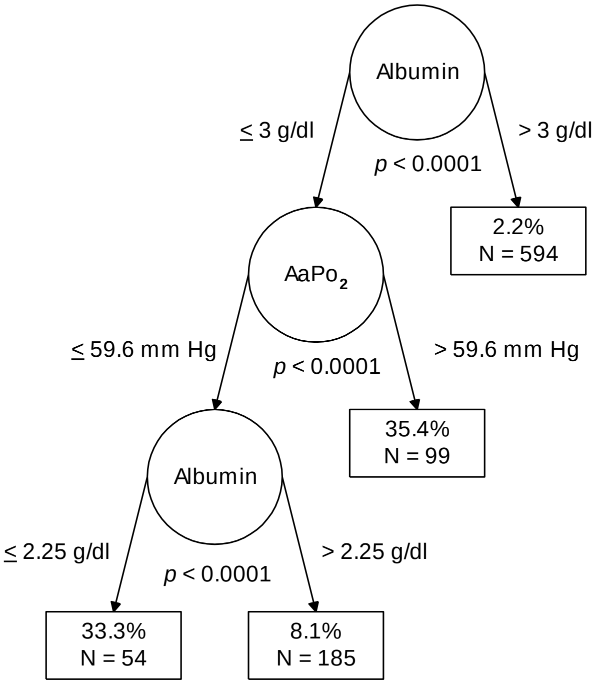

Figure 2: Enumerated CTA Model for Predicting PCP Inpatient Mortality After 1995, Based on Three Attributes

To use the table to stage disease severity for a given patient, simply evaluate fit between patient data and each stage descriptor. Begin at Stage 1, and work sequentially through stages until identifying the descriptor which is true for the data of the patient undergoing staging.

Table 1: Staging Table for Predicting In-Hospital Mortality From PCP, First Model

Stage	Albumin	$AaPo_2$	N	p_{death}	Odds
1	> 3	----	594	0.022	1:44
2	> 2.25	$\leq$ 59.6	185	0.081	1:11
3	$\leq$ 2.25	$\leq$ 59.6	54	0.333	1:2
4	$\leq$ 3	> 59.6	99	0.354	6:11

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

For example, imagine a patient was 54 years of age, male, morbidly obese, with albumin of 2.47 g/dl and $AaPo_2$ of 61.7 mm Hg. Here, age, gender and mass are immaterial to the staging process, because only attributes in the staging table are used in the staging process. Stage 1 does not fit, as the patient's albumin level is less than 3 g/dl. Stage 2 does not fit because the patient's $AaPo_2$ is greater than 59.6 mm Hg. Stage 3 does not fit as the patient's albumin is greater than 2.25 g/dl (when evaluating a descriptor, the first instance of inaccuracy immediately eliminates the Stage from further consideration). Because the staging table has one degree of freedom, Stage 4 must fit: the patient's albumin is less than 3 g/dl, and $AaPo_2$ is greater than 59.6 mm Hg—so Stage 4 indeed fits the data of this hypothetical patient.

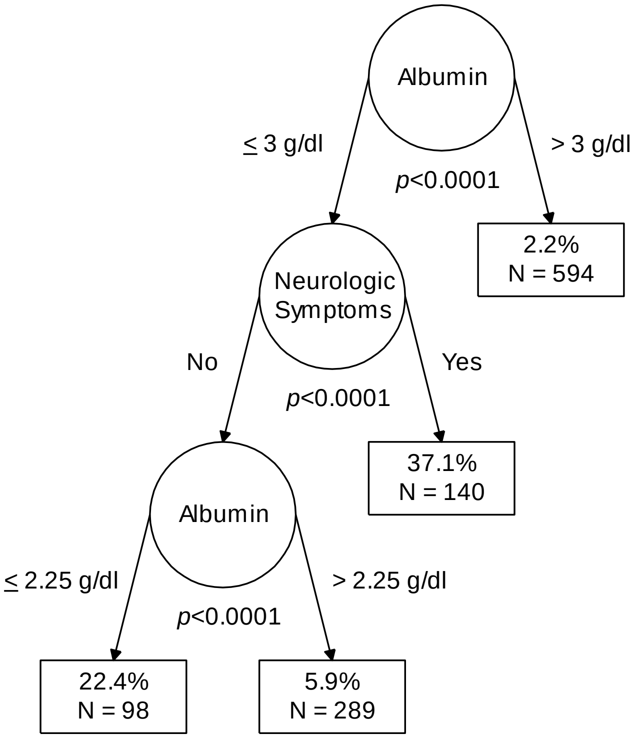

Figure 3: Algorithmic CTA Model Predicting PCP Inpatient Mortality After 1995, Using Attributes From Prior Manual Analysis

Using automated software we next ran automated *algorithmic* CTA (in which the CTA algorithm is performed with optimal parsing but without enumeration), *using all of the attributes employed in original analysis*.[6] A model having three attributes was identified (Figure 3) with 71.2% sensitivity, 83.9% specificity, a *relatively strong* ESS=55.0 (2.5% greater than for the optimized manual model), and efficiency=18.3 ESS units-per-attribute (2.4% greater than for the optimized manual model). Analysis was completed in 85 CPU seconds. The corresponding staging table is presented in Table 2.

Table 2: Staging Table for Predicting In-Hospital Mortality From PCP, Second Model

Stage	Albumin	Neurologic Symptoms	N	p_{death}	Odds
1	> 3	--------	594	0.022	1:44
2	> 2.25	No	289	0.059	1:16
3	≤ 2.25	No	98	0.224	2:7
4	≤ 3	Yes	140	0.371	3:5

An enumerated analysis was conducted next, and a CTA model emerged which yielded a relatively strong effect (ESS=61.4). However, the model included six attributes (two repeated twice), and another attribute which involved a parse. The added complexity, 100% increase in number of attributes employed in exchange for a 11.6% gain in ESS, and 44.1% decrease in efficiency associated with use of the enumerated model, argued in favor of adopting the algorithmic model in this application.

Discussion

Because of inherent importance (having already been judged worthy of publication), and to assemble a literature which may eventually be tapped to assess the magnitude of the boosted

Copyright 2010 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

ESS offered by these methods in real-world applications, all published CTA models derived manually should *minimally* be optimized using UniODA to return maximum ESS, and the pruned models should be published, as is true presently. Of course, all manually derived CTA models should be pruned to maximize ESS prior to consideration.[4] However, current state-of-the-art methodology for achieving maximum ESS involves conducting automated enumerated CTA, which is the optimal choice.

References

[1]Soltysik RC, Yarnold PR. Introduction to automated CTA. *Optimal Data Analysis* In press.

[2]Yarnold PR, Soltysik RC. *Optimal data analysis: a guidebook with software for Windows*. APA Books, Washington, DC, 2005.

[3]Yarnold PR. Discriminating geriatric and nongeriatric patients using functional status information: an example of classification tree analysis via UniODA. *Educational and Psychological Measurement* 1996, 56:656-667.

[4]Yarnold PR, Soltysik RC. Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis* In press.

[5]Yarnold PR, Soltysik RC, Bennett CL. Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: an example of hierarchically optimal classification tree analysis. *Statistics in Medicine* 1997, 16: 1451-1463.

[6]Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, *et al*. A new preadmission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine* 2000, 161:1081-1086.

Author Notes

Mail correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans St., Suite 330, San Diego, CA 92106. Send E-mail to: Journal@OptimalDataAnalysis.com.