Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010), 38-40

2155-0182/10/$3.00

Unconstrained Covariate Adjustment in CTA

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

In traditional statistical covariate analysis it is common practice to force entry of the covariate into the model first, to eliminate the effect of the covariate (i.e., “equate the groups”) on the dependent measure. In contrast, in CTA the covariate is treated as an ordinary attribute which must compete with other eligible attributes for selection into the model based on operator-specified options. This paper illustrates optimal covariate analysis using an application involving predicting patient in-hospital mortality via CTA.

A study of 1,641 patients hospitalized for *Pneumocystis cariini* pneumonia (PCP) used logistic regression analysis to model in-hospital mortality: after forcing a measure of severity-of-illness into the model first, PCP prophylaxis was the only attribute significantly associated with lower hospital survival.[1] During development of an enumerated model involving only these two attributes, a non-pruned[2] CTA model was identified which is analogous to the logistic regression analysis, in that both models *initially adjusted* for severity of illness. CTA analyses were performed using automated software with a minimum endpoint denominator of N=25 to ensure sufficient statistical power.[3] The optimal solution involved one parse of the root attribute (i.e., the first and second attributes entering the CTA model were both PCP severity-of-illness), so the model has three emanating branches (see Figure 1).

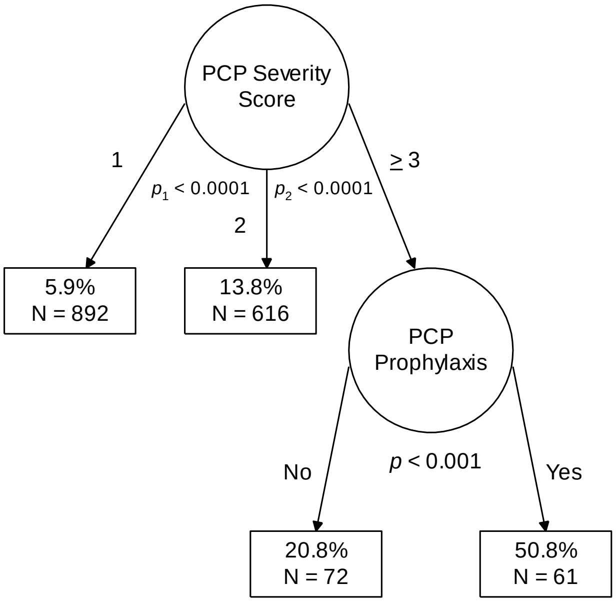

Figure 1: Algorithmic CTA Model Predicting In-Hospital Mortality, Covariate Entered First

Consistent with findings using logistic regression, this CTA model returned weak gain versus chance in predicting mortality: 97.9% of

1,457 living and 16.8% of 184 deceased patients were correctly classified: ESS=14.8, efficiency= 14.8/2 or 7.4 ESS units-per-attribute. Though the CTA *model* is weak, the right-most endpoint indicates that the combination of a PCP severity score of three or greater, and PCP prophylaxis, predicted nearly 51% mortality for 61 patients. Thus, for applications in which it is important to identify particularly vulnerable strata, a variety of different CTA models should be examined in hopes of discovering one or more of such fruitful branches (i.e., combinations).

In contrast, as illustrated in Figure 2, the enumerated CTA model obtained using the same two attributes has robust endpoint denominators; correctly classified 67.9% of the 1,457 living and 61.4% of the 184 dead patients; and obtained moderate strength (ESS=29.4) and efficiency (9.8 ESS units-per-attribute).

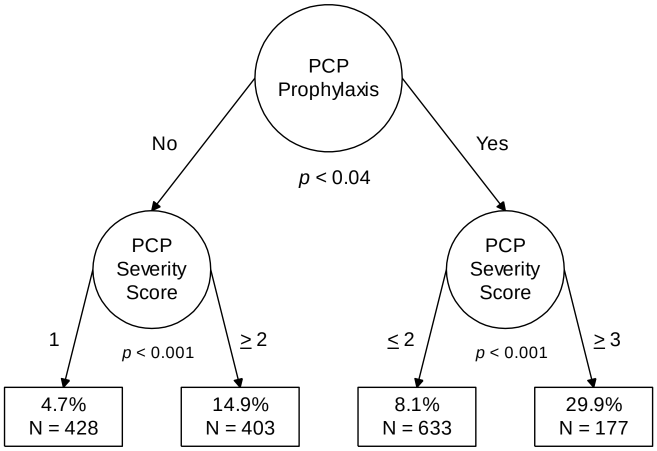

Figure 2: Enumerated CTA Model Predicting In-Hospital Mortality: Covariate Unconstrained

Table 1 gives the staging table for the enumerated CTA model, used for predicting in-hospital mortality from PCP. Table rows are model endpoints reorganized in increasing order of percent of class 1 (“dead”) membership. Stage is an *ordinal index* indicating increasing severity of illness, and p_{death} is a *continuous index* of disease severity. The 1st and 4th strata reflect a 6.4-fold difference in likelihood of dying in-hospital: compared to Stage 1, p_{death} is approximately two times higher in Stage 2, three times higher in Stage 3, and six times higher in Stage 4.

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

	PCP	Severity			
Stage	Prophylaxis	Score	N	p_{death}	Odds
1	No	1	428	0.047	1:20
2	Yes	≤ 2	633	0.081	1:11
3	No	≥ 2	403	0.149	1:6
4	Yes	≥ 3	177	0.299	3:7

Although identical attributes were used by the two CTA models and the original linear logistic regression analysis, the attributes were arranged in different geometries in the different models. Of course, an analyst’s imposition of attribute entry or sequence order in CTA, or any chained optimal analysis, should be performed on the basis of theory, that is, to directly address *a priori* hypotheses.[4] However, the present case clearly indicates the need for caution regarding unchecked rigid adherence to methodological traditions which may actually impede progress achieved using emerging and new technologies. Automated CTA software makes the comparative analysis of multiple theoretical perspectives feasible for most applications: challenging and defeating unfruitful traditions ought to make for interesting, if not exciting research.

References

[1]Curtis JR, Yarnold PR, Schwartz DN, Weinstein RA, Bennett CL (2000). Improvements in outcomes of acute respiratory failure for patients with human immunodeficiency virus-related *Pneumocystis carinii* pneumonia. *American*

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

Journal of Respiratory and Critical Care Medicine, *162*, 393-398.

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

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

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

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.