Optimal Data Analysis
Vol. 8 (January 30, 2019), 38-47

2155-0182/10/$3.00

Optimizing Suboptimal Classification Trees: S-PLUS® Propensity Score Model for Adjusted Comparison of Hospitalized *vs*. Ambulatory Patients with Community-Acquired Pneumonia

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Pruning to maximize model accuracy (requiring simple hand computation) is applied to a classification tree model developed via S-PLUS to create propensity scores to improve causal inference in comparing hospitalized *vs*. ambulatory patients with community-acquired pneumonia. Research reported herein constitutes a thought-provoking example of a striking misalliance between forward analytic thinking and vestige statistical tools—a condition that dominates the empirical literature *today*. Modifications of ubiquitous methodological practices are suggested.

A methodologically sophisticated study conducted longer than two decades ago[1] sought "…to identify a subgroup of patients with community-acquired pneumonia (CAP) who could be safely treated on an ambulatory basis" (p. AS56). This nonrandomized study involved comparing medical outcomes of convenience samples of 265 hospitalized *vs*. 482 ambulatory patients with CAP. Pretreatment differences between hospitalized *vs*. ambulatory patients were controlled by propensity score adjustment. Seven propensity score strata having associated hospitalization probabilities ranging from 6.5% to 76.5% were identified vis-à-vis the S-PLUS® classification tree algorithm.[2] Analysis revealed: "Statistically significant pretreatment imbalances favoring the outpatients were found for 29 of 44 baseline variables considered; after stratification on the propensity score, only 13 of the 29 imbalances remained statistically significant at the 0.05 level. Post hoc stratification on the estimated propensity score consistently reduced, but did not completely eliminate, systematic baseline differences between ambulatory and hospitalized patients with CAP" (p. AS56).

Figure 1 presents a schematic illustration of the seven propensity score strata model that the S-PLUS algorithm identified, and Table 1 summarizes the classification performance (accuracy) which was achieved when using this six-attribute model to classify all 747 patients in total sample "training" analysis.

2155-0182/10/$3.00

Figure 1: Original S-PLUS® Classification Tree Propensity Score Model[1]

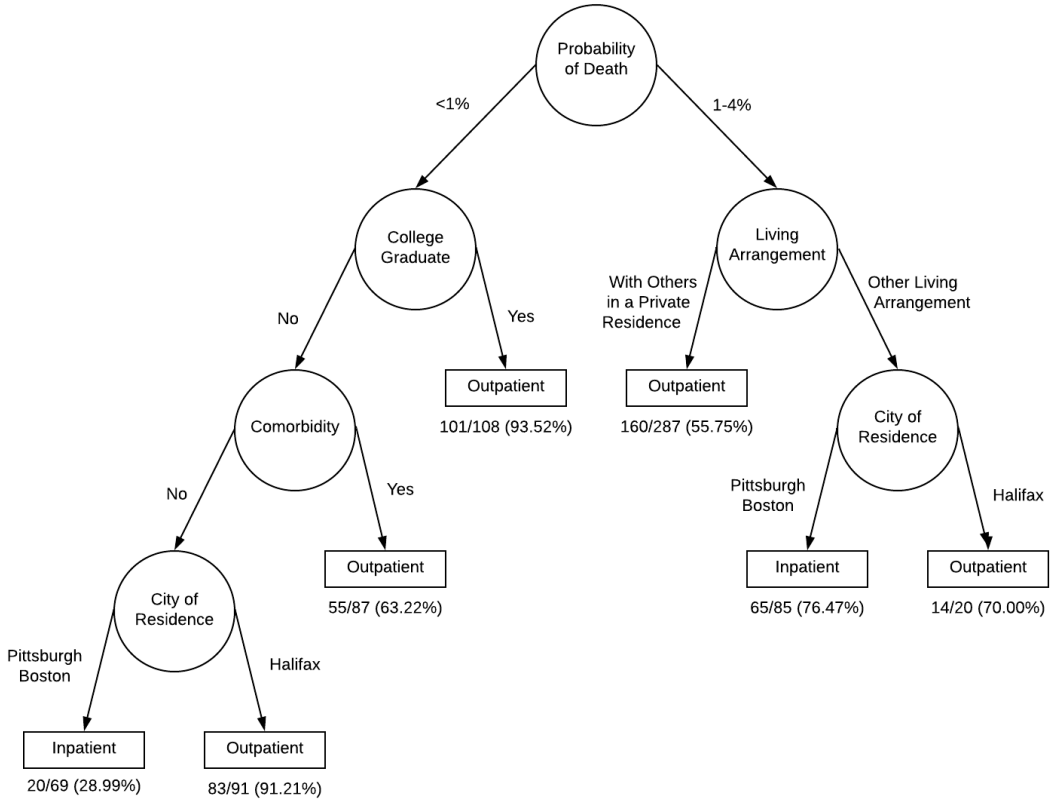


Table 1: Confusion Table for Original S-PLUS Classification Tree Model Shown in Figure 1 (*Sens*=Sensitivity; *PV*=Predictive Value)

	Predicted		
Actual	Inpatient	Outpatient	*Sens*
Inpatient	85	180	32.08
Outpatient	69	413	85.68
PV	55.19	69.65	

Sensitivity findings obtained in training analysis reveal this model correctly classified 32.08% of 265 inpatients: this level of accuracy compares poorly *vs*. chance—for which, defined as a uniform random number, 50% accuracy is expected.[3] In contrast, 85.68% training accuracy for 482 outpatients compares quite favorably *vs*. chance. The *E*ffect *S*trength for *S*ensitivity or ESS index, a transformation of mean sensitivity over class categories, is used to summarize the model omnibus (overall) classification accuracy after adjusting for ("removing") performance expected by chance: ESS=0 is the accuracy expected by chance; ESS=100 is perfect accuracy; and ESS<0 is accuracy worse than expected by chance.[4] Based on Monte Carlo research[5] and applied research occurring over three decades, the rule-of-thumb which is used to qualitatively summarize *effect strength after adjusting for chance* is: ESS<25 is a relatively weak effect; ESS<50 a moderate effect; ESS<75 a relatively strong effect; and ESS≥75 indicates increasingly strong classes of effect size.[6-8] For the model in Figure 1, ESS=17.76—a *relatively weak* effect.

Predictive value findings obtained in training analysis reveal the model to be correct 55.19% of the time it makes a point prediction that a specific observation is an inpatient (154 such point predictions were made), and 69.65%

2155-0182/10/$3.00

when predicting an observation is an outpatient (593 such predictions). *E*ffect *S*trength for *P*redictive Value or ESP index, a transformation of mean PV over class categories, summarizes the model omnibus chance-adjusted PV *for the application*: unlike sensitivity, model PV varies over base rate and is estimated for different base rates.[9] Qualitative strength of the ESP index is determined as for ESS: here ESP=24.84, which is marginally lower than the criterion used to establish the lower bound for an effect having moderate strength. In novometric theory, 95% exact discrete confidence intervals are obtained for the model and for chance (for all measures of performance): overlap of CIs indicates the absence (lack) of statistical significance.[10]

ESS and ESP indices assess translational chance-adjusted accuracy obtained by the model when it is used in actual application. Considered from an applied perspective, a "fully-loaded" model which explicitly maximizes empirically achievable accuracy (ESS) offers the most information available regarding alternative pathways toward and away from the outcome.

However, evaluated from a theoretical perspective such models are considered over-fit: the sought-after model reflects both explanatory *power* (strongest possible ESS) and *parsimony* (fewest possible outcome strata). Theoretical quality of an empirical model is defined in terms of the discrepancy (distance) between achieved *vs*. corresponding perfect model[11], and is quantified using the D (distance) statistic which norms ESS for parsimony: smaller values of D indicate better combinations of accuracy and parsimony, and D=0 indicates a perfect model (number of strata is a function of measurement granularity and distributions over attributes).[12]

For the fully-grown S-PLUS tree model, ESS normed for parsimony is D_{ESS}= 32.41, so 32.41 additional strata having equivalent mean ESS are needed to obtain a "perfect" model (D_{PV}=21.18). Every fully-grown tree model requires optimal pruning to explicitly maximize predictive accuracy normed *vs*. chance.[13]

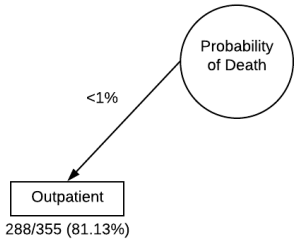
Optimal Pruning to Maximize Model ESS

The optimal pruning algorithm consists of two straightforward steps which are simple to apply by hand. Optimal pruning is the *only* way to *ensure* that a classification tree model explicitly maximizes (weighted) ESS for the sample—no matter what algorithm was employed in development.[13] Optimal pruning was previously illustrated for CART and CTA tree models.[14-16]

Step One of optimal pruning requires identifying all sub-branches of every emanating branch. Imagine a left-hand branch having three nodes: A (root), B (middle attribute), and C (end of branch). There are two nested sub-branches: one involving only nodes A and B (C collapsed into B), the other involving only node A (C and B collapsed into A). Herein the *left* branch with *three* nodes (A, B, C) is known as "L3"; the trimmed *left* branch having *two* nodes (A, C collapsed into B) as "L2"; and the trimmed *left* branch with only *one* node (C and B collapsed into A) as "L1". Imagine also the hypothetical tree model has a right-hand branch having two nodes: A (the sides share the root attribute) and D (end of the branch). The *right* branch having *two* attributes (A, D) is called "R2", and the trimmed *right* branch having *one* attribute (D collapsed into A) is called "R1".

Step One of optimal pruning is executed presently as shown in Figures 2A through 3C.

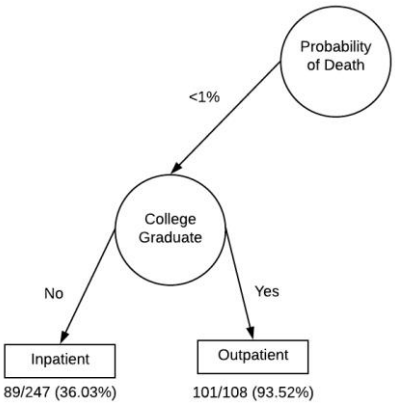
Figure 2A: L1 Sub-Branch and Confusion Table



	Predicted	
Actual	Inpatient	Outpatient
Inpatient	0	67
Outpatient	0	288

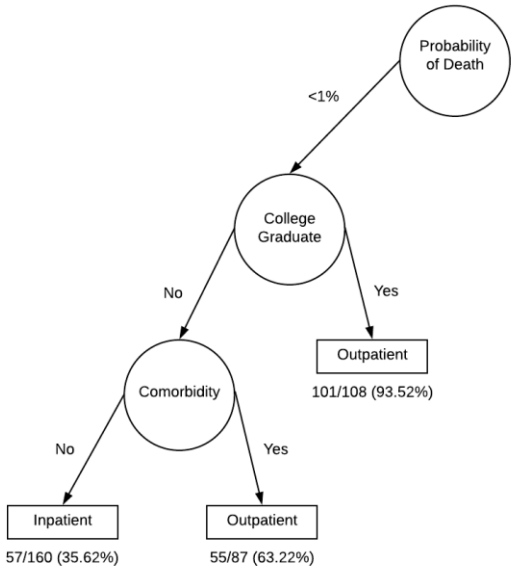
2155-0182/10/$3.00

Figure 2B: L2 Sub-Branch and Confusion Table



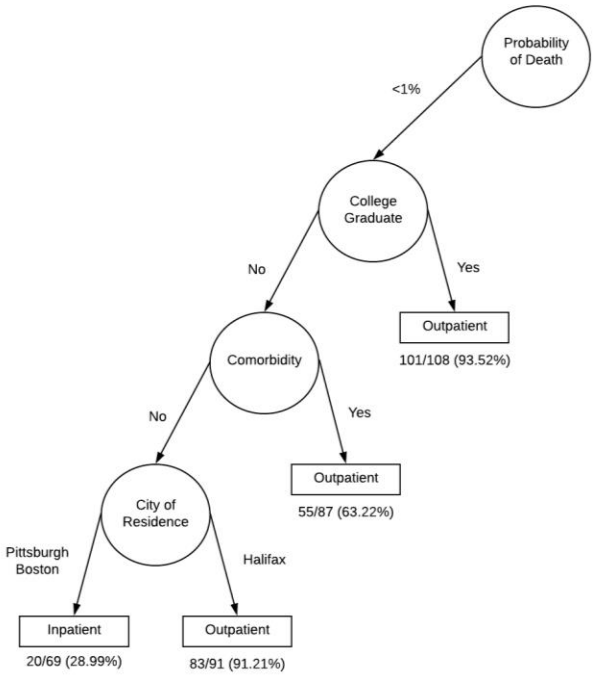
	Predicted	
Actual	Inpatient	Outpatient
Inpatient	89	7
Outpatient	158	101

Figure 2C: L3 Sub-Branch and Confusion Table



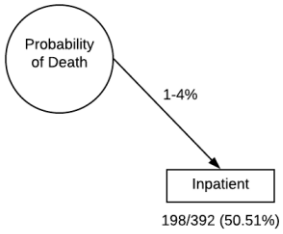
	Predicted	
Actual	Inpatient	Outpatient
Inpatient	57	39
Outpatient	103	156

Figure 2D: L4 Sub-Branch and Confusion Table



	Predicted	
Actual	Inpatient	Outpatient
Inpatient	20	47
Outpatient	49	239

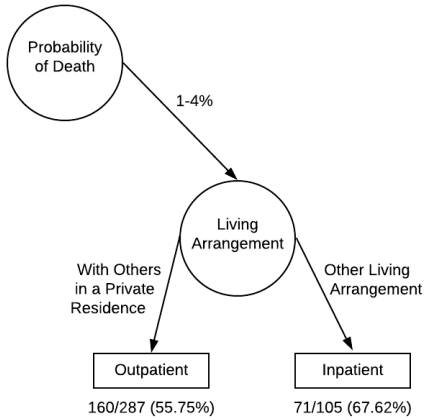
Figure 3A: R1 Sub-Branch and Confusion Table



	Predicted	
Actual	Inpatient	Outpatient
Inpatient	198	0
Outpatient	194	0

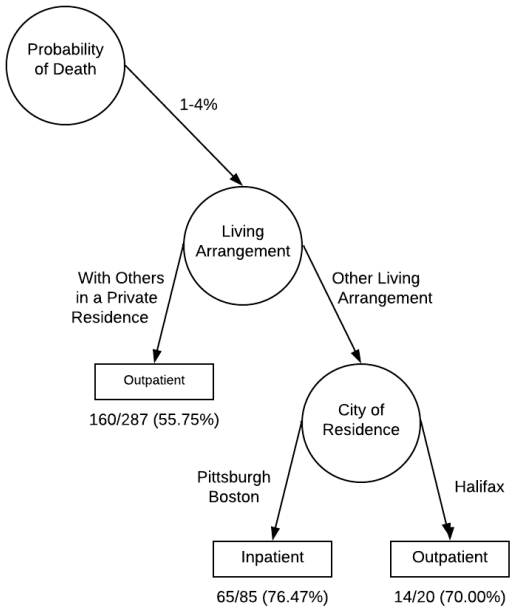
2155-0182/10/$3.00

Figure 3B: R2 Sub-Branch and Confusion Table



	Predicted	
Actual	Inpatient	Outpatient
Inpatient	71	127
Outpatient	34	160

Figure 3C: R3 Sub-Branch and Confusion Table



	Predicted	
Actual	Inpatient	Outpatient
Inpatient	65	133
Outpatient	20	174

Step Two of the algorithm requires creating a confusion table (rows indicate actual class category, columns indicate class category predicted for the observation by the model) for every unique combination of left and right sub-branch. Presently there are 12 unique combinations: {L1-R1, L1-R2, L1-R3}, {L2-R1, L2-R2, L2-R3}, {L3-R1, L3-R2, L3-R3} and {L4-R1, L4-R2, L4-R3}. Table 1 gives integrated confusion tables and associated ESS and D statistics for all unique combinations of left and right branches (red fond indicates strongest effects).

Table 1: Classification Results for Every Combination of Left (L1-L4) and Right (R1-R3) Sub-Branch

Model	Confusion Table	
L1-R1	Predicted	
Actual	Inpatient	Outpatient
Inpatient	198	67
Outpatient	194	288
	ESS=34.5, D=3.80	

Model	Confusion Table	
L1-R2	Predicted	
Actual	Inpatient	Outpatient
Inpatient	71	194
Outpatient	34	448
	ESS=19.7, D=12.2	

Model	Confusion Table	
L1-R3	Predicted	
Actual	Inpatient	Outpatient
Inpatient	65	200
Outpatient	20	462
	ESS=20.4, D=15.6	

Model	Confusion Table	
L2-R1	Predicted	
Actual	Inpatient	Outpatient
Inpatient	287	7
Outpatient	352	101
	ESS=19.9, D=12.1	

2155-0182/10/$3.00

Model	Confusion Table	
L2-R2	Predicted	
Actual	Inpatient	Outpatient
Inpatient	160	134
Outpatient	192	261
	ESS=12.0, D=29.2	

Model	Confusion Table	
L2-R3	Predicted	
Actual	Inpatient	Outpatient
Inpatient	154	140
Outpatient	178	275
	ESS=13.1, D=33.2	

Model	Confusion Table	
L3-R1	Predicted	
Actual	Inpatient	Outpatient
Inpatient	255	39
Outpatient	297	156
	ESS=21.2, D=14.9	

Model	Confusion Table	
L3-R2	Predicted	
Actual	Inpatient	Outpatient
Inpatient	128	166
Outpatient	137	316
	ESS=13.3, D=32.6	

Model	Confusion Table	
L3-R3	Predicted	
Actual	Inpatient	Outpatient
Inpatient	122	172
Outpatient	123	330
	ESS=14.4, D=35.8	

Model	Confusion Table	
L4-R1	Predicted	
Actual	Inpatient	Outpatient
Inpatient	218	47
Outpatient	243	239
	ESS=31.8, D=10.7	

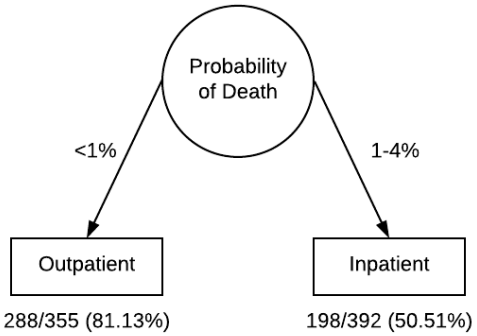
Model	Confusion Table	
L4-R2	Predicted	
Actual	Inpatient	Outpatient
Inpatient	91	174
Outpatient	83	399
	ESS=17.1, D=29.1	

Model	Confusion Table	
L4-R3	Predicted	
Actual	Inpatient	Outpatient
Inpatient	85	180
Outpatient	69	413
	ESS=17.8, D=32.4	

The optimized model is the combination of left and right sub-branches with associated confusion table yielding maximum ESS. Seen in Table 1, the combination L1-R1 (see Figure 4) has greatest mean sensitivity (67.23%), yielding ESS=34.5—a moderate effect. However, L1-R1 is not only the best *empirical* model—by virtue of having lowest value of D among all models, L1-R1 is also closest to representing a *theoretically* optimal solution.

Figure 4

Optimized S-PLUS Classification Tree Propensity Score Model



In summary, a classification tree propensity score model developed using a suboptimal[17] algorithm yielded relatively weak ESS; the tree model was pruned to explicitly obtain maximum

2155-0182/10/$3.00

possible ESS; and a single-attribute model emerged which achieved a moderate effect in training analysis. The attribute was probability of death (estimated using a logistic regression model obtained for another sample) treated as a categorical attribute defined (<1% *vs*. 1-4%) on the basis of expert (authors') opinion.

Thus, in essence, all analyses in this and in the original[1] (prior) study—everything but the logistic regression analysis reported two studies ago—was superfluous?! What happened!?

Passé Methodological Factors Limiting Model Accuracy and Threatening Internal Validity: Methodological Recommendations

For expositional clarity and efficiency, different components/aspects of the fully grown S-PLUS classification tree model and the recommended modifications are separately discussed below.

Consideration begins with the root attribute of the S-PLUS classification tree model (Figure 1), described as assessing "estimated probability of death." It is important to clarify that the study[1] estimated probability of 60-day in-hospital mortality for every patient using a logistic regression model developed for a separate sample of patients with CAP. Patients with a predicted probability of death greater than 4% were eliminated because investigators believed few clinicians would treat such patients on an ambulatory basis (p. AS58). After discussion investigators selected 1% to use as the threshold value defining estimated likelihood of mortality: each patient was scored using a binary indicator of whether estimated probability of death was <1% (category 0) or 1-4% (category 1).

Two points of order are proper. First, an accurate label is needed for every attribute. For example, for the root attribute, "prior-logistic-regression-model-based estimated probability of 60-day in-hospital mortality dichotomized into categories <1% or 1-4%" (or something more "catchy"). Second, the opinion that few patients with >4% estimated probability of death would likely be treated on an ambulatory basis requires evaluation as an *a priori* hypothesis.[18] These points further call into question the empirical adequacy of the dichotomized attribute.[19]

First, this attribute is *not* qualitative but rather it reflects the lowest level of granularity possible for an ordinal (ordered) attribute: the "category" of <1% clearly reflects a lower value (magnitude) than does the "category" of 1-4%.[20]

Second, a statistically unmotivated dichotomous split such as <1% *vs*. 1-4%—indeed, *any* distribution (splitting) or agglomeration (combining) of ordinal data may be incorrectly placed and/or insufficiently granular to identify the model yielding strongest results.[21] Incorrect clustering and separating of qualitative categories of multicategorical attributes similarly has adverse consequences.[22,23] Of course, for some applications involving testing *a priori* hypotheses or performing sensitivity analyses, one-tailed evaluation is integral in planned analyses optimizing one[24] or more[25] objective functions.

Third, available precision of the attribute is imbalanced below *vs*. above threshold. Below threshold the estimated probability of death (or p_{death}) for Class=0 members can vary relatively dramatically. For example a Class=0 observation with estimated p_{death}=0.00999 is 99.9 times more likely to die than another Class=0 observation having estimated p_{death}= 0.0001. In contrast, above threshold the most p_{death} can vary for two Class=1 observations is 0.004/0.001, for a maximum-possible 4-fold difference in p_{death}.

Fourth, combining data for observations drawn from heterogeneous populations—such as patient groups which differ by two orders of magnitude in estimated likelihood of 60-day in-hospital mortality, for example—into a single group can create Simpson's Paradox in which analysis of combined data can produce findings that differ from findings obtained by the same analysis conducted separately by group.[26,27]

And, while it is axiomatic that obtaining reproducible results is crucial in programmatic research, thus potential cross-generalizability of a model should always be estimated—it isn't.

2155-0182/10/$3.00

Fortuitously, the use of globally-optimal (GO) *novometric* (Latin: New Measurement) methods[3,6] eliminate issues identified for the root variable. First, redefine p_{death} on the basis of a GO-CTA model having stable classification accuracy in one-sample "leave-one-out" (LOO) jackknife analysis (to maximize cross-generalizability), and subsequently treat p_{death} as an ordinal attribute. Enumerated-optimal CTA models are available for *Pneumocystis carinii* pneumonia[28,29], CAP[30,31] and inhalational anthrax infection[32], and provide excellent starting frames of reference in this analysis. Second, use GO-CTA to identify all unique CTA models in the sample which vary by parsimony (number of strata) and statistical power (minimum strata sample size), and identify the model having greatest translational value (ESS) and theoretical quality (D). The stated objective of the original study[1]—to "identify *a subgroup* of patients" with CAP who could be safely treated on an ambulatory basis—is modest in light of the power of optimal methods to identify all existing subgroups at all levels of statistically viable model granularity.

Consideration moves to the second attribute of the S-PLUS classification tree model, a binary indicator called "College Graduate" seen beneath and to the left of the root attribute. This attribute is clearly ordinal, as "college" reflects a greater amount of education than "no college". It may prove useful to further graduate the response scale by adding High School as a low-end marker of educational attainment, and Post Graduate as a high-end. Ultimately it may prove to be most useful to identify and measure specific aspect(s) of "education" which serve patients and other decision-makers in assessing appropriate placement (inpatient *vs*. outpatient) for each individual.

Consider next the binary indicator called "Comorbidity" having response options Yes *vs*. No, seen beneath and left of the College Graduate attribute. It seems unreasonable to equate the "value" of comorbidities to predict inpatient status: neurological comorbidities are likely more influential in this regard than, for example, is hypertension. Accordingly, it is a much better idea to dummy-code the various comorbidities in a single multicategorical attribute, and let the CTA algorithm determine the optimal structure underlying comorbidities if the attribute is used in the model.[33]

The next attribute to consider, to the right of and below the root attribute, is called Living Arrangement. This attribute reflects a 2x2 factorial design: the orthogonal 2-category (Yes, No) main effects are "Live with Others" and "In a Private Residence". In the model the left path from this attribute is the interaction of these main effects, and the right path ("other") is thus both main effects. The ESS of this model would be most improved by adding an attribute to the left branch from Living Arrangement—which presently misclassifies 127 inpatients. As for College Education, it may prove most useful to identify and measure specific aspect(s) of "lives with others" and/or of "in a private residence" which accurately predict inpatient *vs*. outpatient placement of individuals.

The final attribute, a multicategorical indicator of city of residence, may not load in a GO-CTA model—typically involving fewer attributes than used in Figure 1, because the attribute city of residence may not produce higher ESS than other attributes.

References

[1]Stone RA, Obrosky DS, Singer DE, Kapoor WN, Fine MJ, and The Pneumonia Patient Outcomes Research Team (PORT) Investigators (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired Pneumonia. *Medical Care*, *4*, AS56-AS66.

[2]Statistical Sciences, Inc. (1993). *S-PLUS for Windows user's manual, Version 3.1*. Seattle, WA: Statistical Sciences, Inc.

2155-0182/10/$3.00

[3]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis, 6,* 26-42.

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

[5]Yarnold PR, Soltysik RC (2010). Precision and convergence of Monte Carlo Estimation of two-category UniODA two-tailed *p*. *Optimal Data Analysis*, *1*, 43-45.

[6]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[7]Bryant FB (2010). The Loyola experience (1993-2009): Optimal Data Analysis in the Department of Psychology. *Optimal Data Analysis, 1,* 4-9.

[8]Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, *1,* 10-22.

[9]Ostrander R, Weinfurt KP, Yarnold PR, August G (1998). Diagnosing attention deficit disorders using the BASC and the CBCL: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology, 66,* 660-672.

[10]Yarnold PR (2018). Comparing exact discrete 95% CIs for model *vs*. chance ESS to evaluate statistical significance. *Optimal Data Analysis, 7,* 82-84.

[11]Yarnold PR (2019). The structure of *perfect* optimal models with a two-category class variable and four or fewer endpoints. *Optimal Data Analysis, 8,* 21-25.

[12]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis, 5,* 171-174.

[13]Yarnold PR, Soltysik RC (2010). Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis, 1,* 23-29.

[14]Yarnold PR (2016). Pruning CTA models to maximize PAC. *Optimal Data Analysis, 5,* 58-61.

[15]Yarnold PR (2019). Maximizing classification accuracy of CART® recursive partitioning tree models using optimal pruning. *Optimal Data Analysis, 8,* 26-29.

[16]Yarnold PR (2019). Maximizing the accuracy of a CART tree model predicting missing data. *Optimal Data Analysis, 8,* 33-37.

[17]Yarnold PR (2014). "A statistical guide for the ethically perplexed" (Chapter 4, Panter & Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge, 2011): Clarifying disorientation regarding the etiology and meaning of the term *Optimal* as used in the Optimal Data Analysis (ODA) paradigm. *Optimal Data Analysis, 3,* 30-31.

[18]Yarnold PR (2014). Increasing the validity and reproducibility of scientific findings. *Optimal Data Analysis, 3,* 107-109.

[19]Yarnold PR (2018). Minimize usage of binary measurement scales in rigorous classical research. *Optimal Data Analysis, 7,* 3-9

[20]Yarnold PR (2010). UniODA *vs*. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis, 1,* 62-65.

[21]Yarnold PR (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis, 3,* 19.

[22]Yarnold PR (2010). Aggregated *vs*. referenced categorical attributes in UniODA and CTA. *Optimal Data Analysis, 1,* 46-49.

2155-0182/10/$3.00

[23]Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, *2*, 177-190.

[24]Harvey RL, Roth EJ, Yarnold PR, Durham JR, Green D (1996). Deep vein thrombosis in stroke: The use of plasma D-dimer level as a screening test in the rehabilitation setting. *Stroke, 27,* 1516-1520.

[25]Yarnold PR (2019). Growing classification tree models on the basis of *a priori* performance criteria. *Optimal Data Analysis*, *8,* 30-32.

[26]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, *56,* 430-442.

[27]Bryant FB, Siegel EKB (2010). Junk science, test validity, and the Uniform Guidelines for Personnel Selection Procedures: The case of *Melendez v. Illinois Bell*. *Optimal Data Analysis, 1,* 176-198.

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

[29]Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, Chmiel JS, Sipler AM, Chan C, Goetz MB, Schwartz D, Bennett CL (2000). 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*, *161*, 1081-1086.

[30]Arozullah AM, Parada J, Bennett CL, Deloria-Knoll M, Chmiel JS, Phan L, Yarnold PR (2003). A rapid staging system for predicting mortality from HIV-associated community-acquired pneumonia. *Chest*, *123*, 1151-1160.

[31]Kyriacou DM, Yarnold PR, Soltysik RC, Wunderink RG, Schmitt BP, Parada JP, Adams JG (2008). Derivation of a triage algorithm for chest radiography of community-acquired pneumonia in the emergency department. *Academic Emergency Medicine*, *15*, 40-44.

[32]Kyriacou DN, Yarnold PR, Stein AC, Schmitt BP, Soltysik RC, Nelson RR, Frerichs RR, Noskin GA, Belknap SB, Bennett CL (2007). Discriminating inhalational anthrax from community-acquired pneumonia using chest radiograph findings and a clinical algorithm. *Chest*, *131,* 489-495.

[33]Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, *2*, 177-190.

Author Notes

No conflict of interest was reported.