2015

Obtaining an Enumerated CTA Model via Automated CTA Software

Paul R. Yarnold
Optimal Data Analysis LLC

Fred B. Bryant
Loyola University Chicago, fbryant@luc.edu

Optimal Data Analysis
Vol. 4 (May 14, 2015), 54-61

2155-0182/10/$3.00

Obtaining an Enumerated CTA Model via Automated CTA Software

Paul R. Yarnold, Ph.D. and Fred B. Bryant, Ph.D.
Optimal Data Analysis, LLC Loyola University Chicago

The use of automated CTA software to obtain an enumerated optimal (maximum-accuracy) classification tree analysis (EO-CTA) model is demonstrated and the resulting model is compared with a HO-CTA model developed using the same data.

The development of methodology for obtaining hierarchically optimal classification tree analysis (HO-CTA) models using either UniODA[1] or MegaODA[2-4] statistical software yielded models in numerous disciplines that were more accurate, parsimonious and theoretically apropos than complementary linear models developed using legacy general linear model and maximum-likelihood paradigms.[5] However, manual construction of a maximum-accuracy HO-CTA model is a complex and an analysis-intensive enterprise.[5] This requirement for rigorous computation motivated the development of automated statistical software capable of identifying HO-CTA models, as well as previously inconceivable enumerated optimal classification tree analysis (EO-CTA) models.[6] Whereas HO-CTA models begin with the attribute yielding highest ESS in the root node or the tree model, EO-CTA models evaluate all combinations of attributes in the top three nodes of the tree model.[6] Availability of this automated CTA software yielded models in numerous disciplines that were more accurate, parsimonious and theoretically apropos than corresponding linear models developed using legacy[7-31] or HO-CTA[32-34] methods. The present article demonstrates how to obtain an EO-CTA model with automated CTA software.[6]

Context of the Exposition

As described in the exposition of the development of an HO-CTA model[5], data for this exposition came from a study investigating factors increasing the likelihood of an ambivalent Emergency Department (ED) patient recommending the ED to others. The study was set in an urban 800 bed university-based level 1 Trauma center with annual census of 48,000 patients.[35] One week post discharge, patients were mailed a survey assessing satisfaction with care received in the ED. The survey elicited ratings of the likelihood of recommending the ED to others, and satisfaction with aspects of administration, nurse, physician, laboratory, and care of family/friends. A total of 2,109 surveys with completed recommendation ratings were returned over a six-month period (17% return rate). Likelihood to recommend ("recom" in the UniODA code) was rated using a five-point Likert-type scale: scores of 3 (fair, N=239) indicate *ambivalence*; and scores of 4 (good, N=584) reflect *likely to recommend*.[63] Analysis

2155-0182/10/$3.00

thus included a total of 823 patients responding with recommendation ratings of 3 or 4.

As was done in the demonstration of the development of the HO-CTA model[6], in this exposition only satisfaction ratings of aspects of care received from nurses were used as potential attributes: n1=courtesy; n2=took the patient's problem seriously; n3=attention; n4= informed patient about treatment; n5=concern for privacy; and n6=technical skill. Satisfaction items were completed using five-point Likert-type scales: scores of 1=very poor satisfaction, 2=poor, 3=fair, 4=good and 5=very good satisfaction. Data file requirements for CTA software are the same as for UniODA software, and are discussed elsewhere.[36]

Determining the Minimum *N* for CTA Model Endpoints

The first step in developing any CTA model is to determine *a priori* the minimum appropriate sample size for any (every) endpoint in the model. Two issues that require consideration in this context include statistical power and cross-sample generalizability.[1] As is detailed in exposition of HO-CTA analysis of the present data, consideration of statistical power and generalizability considerations determined that the minimum endpoint value in this application is 42 observations.[5] In order to enter the EO-CTA model, the attribute with the highest ESS value must meet the criterion for experimentwise statistical significance, and must also have an endpoint with 42 or more observations.

Obtaining the EO-CTA Model

The HO-CTA and EO-CTA models for this application *were both generated* using the following CTA[6] code:

```
OPEN recom.dat;
OUTPUT recom.out;
VARS recom n1 to n6;
CLASS recom;
ATTR n1 to n6;
MISSING all (-9);
MC ITER 10000 CUTOFF .05 STOP 99.9;
PRUNE .05;
ENUMERATE;
MINDENOM 42;
GO;
```

Note that the commands used to operate CTA software are the same as the commands used to operate UniODA and MegaODA software, except for the Monte Carlo simulator and the three following commands.[5] The Monte Carlo (MC) simulator is designed to stop when there is a confidence level of less than 99.9% that $p<0.05$ has been obtained (UniODA and MegaODA have the same capability, but the MC command is parameterized in CTA to speed solution time—this is a less of an issue when conducting UniODA analysis). The PRUNE command specifies Sidak-based experimentwise pruning at the specified Type I error rate (*p*-value).[5] The ENUMERATE command specifies that an enumerated CTA model is sought: eliminating this command obtains a HO-CTA model; expressing this command obtains both an HO-CTA model and an EO-CTA model.[5] Here the identical HO-CTA model manually identified using UniODA[5] was provided in the output of the present automated CTA analysis. The MINDENOM command specifies the minimum *N* allowed in every endpoint of the model.[5] Automated CTA required 4 CPU seconds to conduct the HO-CTA analysis, and an additional 48 CPU seconds to conduct the EO-CTA analysis, when run on a 3 GHz Intel Pentium D microcomputer.

The HO-CTA model that was identified automatically using CTA software, and that was identified mechanically using either UniODA or MegaODA software, is presented as Figure 7 in Yarnold and Bryant[5] (p. 45). Figure 1 presents the EO-CTA model identified presently using automated CTA software. Table 1 presents the

2155-0182/10/$3.00

confusion table for this model applied to the data (note that the sample is reduced to N=748 due to missing data).

Figure 1: EO-CTA Model

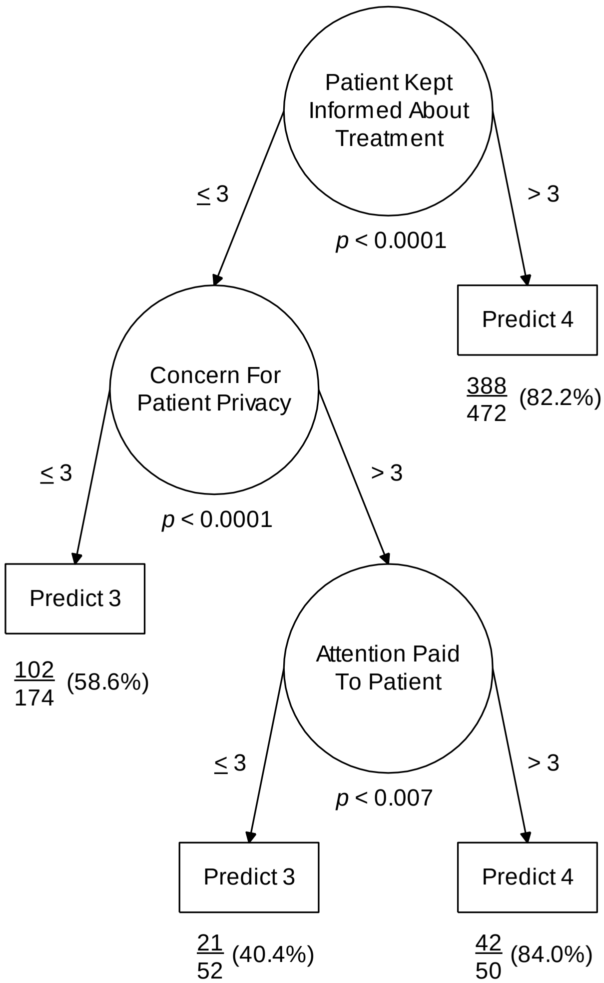


Table 1: Confusion Table for EO-CTA Analysis

		Predicted Recommendation	
		3	4
Actual Recommendation	3	123	92
	4	103	430

As seen, when the model predicted a recommended likelihood score of 3, a total of 103 observations were misclassified; and when the model predicted a recommended likelihood score of 4, a total of 92 observations were misclassified. The sensitivity of this model for class category 3 is 123 / (123 + 92) = 0.572, and the sensitivity of this model for class category 4 is 430 / (430 + 103) = 0.807. The mean sensitivity is thus 0.690, and ESS = [(0.690 – 0.5) / 0.5] x 100% = 37.9.

Developed using this EO-CTA model, Table 2 presents a staging table for predicting the likelihood of a patient recommending the ED to others.[6] Stage is an ordinal index of the likelihood of the patient recommending the ED to others; p_{recom} is a more granular ordered index of the likelihood of the patient recommending the ED to others.

Table 2: Staging Table for Predicting Likelihood of Recommending ED to Others

Stage	Informed About Treatment	Concern for Patient Privacy	Attention Paid To Patient	N	p_{recom}	Odds
1	≤ 3	> 3	≤ 3	52	.404	2:3
2	≤ 3	≤ 3	----	174	.586	3:2
3	> 3	----	----	472	.822	9:2
4	≤ 3	> 3	> 3	50	.840	5:1

Note: p_{recom} = likelihood of recommending ED to others, and Odds = odds of recommending ED to others.

The attribute importance in discrimination (AID) statistic is conceptually similar to the R^2 statistic in regression analysis: both statistics indicate the importance of every attribute in the model with respect to predicting the value of the class variable.[6] The most important attribute is the root node—nurse informed patient about treatment: this attribute was used in predicting class category status of all observations (AID= 100%). The second-most-important attribute was concern for patient privacy, which was in-

2155-0182/10/$3.00

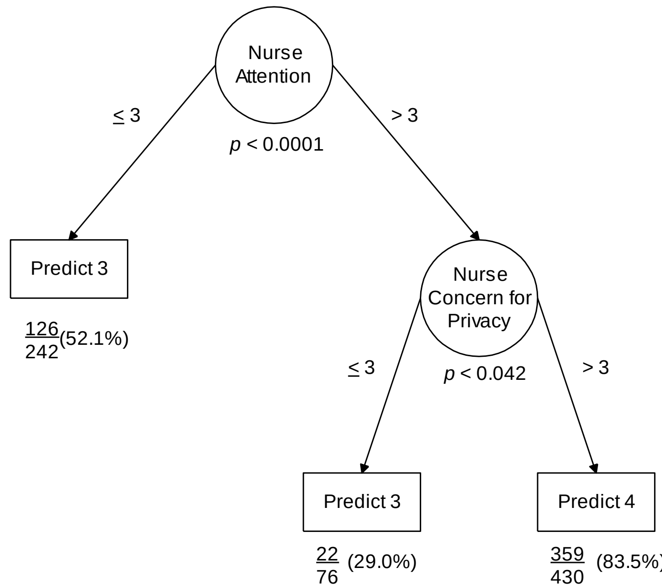
strumental in classification of (174 + 52 + 50) / 748 = 36.9% of the observations. Least important was attention paid to patient: (52 + 50) / 748 = 13.6% of observations.

When considered from a redundancy perspective stages 3 and 4 of this EO-CTA model clearly predict approximately the same proportion of patients likely to recommend the ED to others—albeit for different reasons. However, the concept of redundancy primarily applies to models identifying multiple strata for a single attribute.[37]

The most important substantive revelation of this EO-CTA model is the importance of the nurse keeping the patient informed about treatment: for 472 / 748 = 63.1% of the sample, 4 of 5 patients rating this attribute as good or very good were likely to recommend the ED to others. And, for the remaining 36.9% of the sample rating this attribute as fair or worse, (42 + 21) / 102 = 62% of the patients were likely to recommend the ED to others if the nurse's concern for their privacy was rated as good or very good. These are actionable behaviors that should be emphasized in an effort to maximize positive patient recommendations of the ED.

It is informative to consider the similarities and differences between the three-attribute EO-CTA model constructed in the present analysis (ESS = 37.9; see Figure 1) and the two-attribute HO-CTA model constructed in the earlier analysis[5] (ESS = 35.4; see Figure 2). With respect to *similarities* between the two types of models, the EO-CTA and HO-CTA models include two of the same attributes—namely, concern for patient privacy, and attention paid to patient—each of which has the same optimal cut-point (i.e., 3) in both models. In addition, values > 3 for both attributes produce nearly identical predictive values for the deepest right-hand endpoint in both models (84.0% for the EO-CTA model vs. 83.5% for the HO-CTA model), although this combination of higher values of the two attributes involves very different sample sizes in the two models (i.e., 42/50 in the EO-CTA model vs. 359/430 in the HO-CTA model; or a sample size roughly 8.5 times greater in the HO-CTA model).

Figure 2: Final Pruned Maximum-Accuracy HO-CTA Model[5]



With respect to *difference*s between the two types of models, although the EO-CTA and HO-CTA models include two of the same attributes, these two attributes appear in opposite order in the two models—in the EO-CTA model, concern for patient privacy enters before attention paid to patient, whereas in the HO-CTA model, attention paid to patient enters before concern for patient privacy. Furthermore, in the EO-CTA model, this combination of concern for patient privacy and attention paid to patient is relevant only for patients who were relatively dissatisfied with how well informed they were about their treatment; whereas in the HO-CTA model, this same two-attribute combination (albeit in opposite order of entry) constitutes the full tree model. Thus, for this particular set of attributes, the EO-CTA model qualifies the HO-CTA model by clarifying that the interaction of nurse attention and nurse concern for privacy in predicting likelihood of recommend-

2155-0182/10/$3.00

ing the ED to others is most applicable to patients who are less satisfied with the degree to which the nurse kept them informed about their treatment.

It is also instructive to compare the initial (root) nodes of the EO-CTA and HO-CTA models. The first attribute to enter the EO-CTA model at the root node is patient kept informed about treatment, which predicts a likelihood-of-recommending rating of 4 with 82.2% accuracy. In contrast, the first attribute to enter the HO-CTA model at the root node is nurse attention paid to patient, which predicts a likelihood-of-recommending rating of 4 with 81.5% accuracy. Given that the HO-CTA model always begins with the single strongest predictor at the initial (root) node, one might think that the attribute of patient kept informed would enter the initial (root) node of both the EO-CTA and HO-CTA models.

However, it is not an attribute's predictive accuracy for one or the other levels of the dichotomous class variable, but rather its overall ESS, that determines its entry in the initial (root) node of the HO-CTA model. Computing ESS for the UniODA model using the attribute of patient kept informed about treatment to predict patients' likelihood-of-recommending rating, we find that ESS=29.7. And computing ESS for the UniODA model using the attribute of nurse attention paid to patient to predict patients' likelihood-of-recommending rating, we find that ESS=35.1. Thus, the attribute of nurse attention paid to patient entered the initial (root) node of the HO-CTA model because it has the highest overall ESS of all the attributes in the analysis.

However, in the EO-CTA model, all possible permutations of the attributes being analyzed are enumerated for the first three levels of the model, to find the combination of attributes that maximizes overall classification accuracy for the entire model as a whole. In the present case, entering patient kept informed about treatment at the initial (root) node produced the particular three-attribute combination of predictors that optimizes overall classification accuracy for the integrated model.

References

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

[2]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, *2*, 194-197. URL: http://odajournal.com/2013/11/19/megaoda-large-sample-and-big-data-time-trials-separating-the-chaff/

[3]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the wheat. *Optimal Data Analysis*, *2*, 202-205. URL: http://odajournal.com/2013/11/21/megaoda-large-sample-and-big-data-time-trials-harvesting-the-wheat/

[4]Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, *2*, 220-221. URL: http://odajournal.com/2013/11/27/megaoda-large-sample-and-big-data-time-trials-maximum-velocity-analysis/

[5]Yarnold PR, Bryant FB (2015). Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis*, *4*, 36-53. URL: http://odajournal.com/2015/05/11/obtaining-a-hierarchically-optimal-cta-model-via-unioda-software/

[6]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160. URL: http://odajournal.com/2013/09/19/62/

[7]Arozullah AM, Gordon HS, Yarnold PR, Soltysik RC, Ferreira MR, Wolf MS, Molokie R, Bhoopalam N, Bennett CL (2008). Predictors of prostate cancer stage at presentation. *Journal of General Internal Medicine*, *23*, 376.

2155-0182/10/$3.00

[8]Arozullah AM, Lee SD, Khan T, Kurup S, Ryan J, Bonner M, Soltysik RC, Yarnold PR (2006). The roles of low literacy and social support in predicting the preventability of hospital admission. *Journal of General Internal Medicine*, *21*, 140-145. DOI:10.1111/j.1525-1497.2005.00300.x

[9]Belknap SM, Moore H, Lanzotti SA, Yarnold PR, Getz M, Deitrick DL, Peterson A, Akeson J, Maurer T, Soltysik RC, Storm J (2008). Application of software design principles and debugging methods to an analgesia prescription reduces risk of severe injury from medical use of opioids. *Clinical Pharmacology and Therapeutics*, *84,* 385-392. DOI: 10.1038/clpt.2008.24

[10]Bryant FB, Yarnold PR (2014). Finding joy in the past, present, and future: The relationship between Type A behavior and savoring beliefs among college undergraduates. *Optimal Data Analysis*, *3*, 36-41. URL: http://odajournal.com/2014/04/10/finding-joy-in-the-past-present-and-future-the-relationship-between-type-a-behavior-and-savoring-beliefs-among-college-undergraduates/

[11]Bryant FB, Yarnold PR (2014). Type A behavior, pessimism and optimism among college undergraduates. *Optimal Data Analysis*, *3*, 32-35. URL: http://odajournal.com/2014/04/10/type-a-behavior-pessimism-and-optimism-among-college-undergraduates/

[12]Collinge WC, Kahn J, Walton T, Kozak L, Bauer-Wu S, Fletcher K, Yarnold PR, Soltysik RC (2013). Touch, caring, and cancer: Randomized controlled trial of a multimedia caregiver education program. *Supportive Care in Cancer*, *21*, 1405-1414. DOI: 10.1007/s00520-012-1682-6

[13]Collinge WC, Soltysik RC, Yarnold PR (2010). An internet-based intervention for fibromyalgia self-management: Initial design and alpha test. *Optimal Data Analysis, 1*, 163-175. URL: http://odajournal.com/2013/09/19/an-internet-based-intervention-for-fibromyalgia-self-management-initial-design-and-alpha-test/

[14]Collinge WC, Yarnold PR, Soltysik RC (2013). Fibromyalgia symptom reduction by online behavioral self-monitoring, longitudinal single subject analysis and automated delivery of individualized guidance. *North American Journal of Medical Sciences*, *5*, 546-553. DOI: 10.4103/1947-2714.118920

[15]Grobman WA, Terkildsen MF, Soltysik RC, Yarnold PR (2008). Predicting outcome after emergent cerclage using classification tree analysis. *American Journal of Perinatology*, *25*, 443-448. DOI: 10.1055/s-0028-1083843

[16]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.

[17]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. DOI: 10.1111/j.1553-2712.2007.00011.x

[18]Lavigne JV, LeBailly SA, Gouze KR, Binns HJ, Keller J, Pate L (2010). Predictors and correlates of completing behavioral parent training for the treatment of oppositional defiant disorder in pediatric primary care. *Behavior Therapy*, *41*, 198-211. DOI: 10.1016/j.beth.2009.02.006

[19]Lyons AM, Leon SC, Zaddach C, Luboyeski EJ, Richards M (2009). Predictors of Clinically Significant Sexual Concerns in a Child Welfare Population. *Journal of Child and Adolescent Trauma*, *2*, 28-45. DOI: 10.1080/19361520802675884

2155-0182/10/$3.00

[20]Nebeker JR, Yarnold PR, Soltysik RC, Sauer BC, Sims SA, Samore MH, Rupper RW, Swanson KM, Savitz LA, Shinogle J, Xu W (2007). Developing indicators of inpatient adverse drug events through non-linear analysis using administrative data. *Medical Care*, *45*, S81-S88. URL: http://www.effectivehealthcare.ahrq.gov/repFiles/MedCare/s81.pdf

[21]Sieracki JH, Fuller AK, Leon SC, Jhe Bai G, Bryant FB (2015). The role of race, socioeconomic status, and System of Care services in placement decision-making. *Children and Youth Services Review*, DOI: 10.1016/j.childyouth.2014.12.013

[22]Smith JH, Bryant FB, Njus D, Posavac EJ (2010). Here today, gone tomorrow: Understanding freshman attrition using Person-Environment Fit Theory. *Optimal Data Analysis*, *1*, 101-124. URL: http://odajournal.com/2013/09/19/here-today-gone-tomorrow-understanding-freshman-attrition-using-person-environment-fit-theory/

[23]Snowden J, Leon S, Sieracki J (2008). Predictors of children in foster care being adopted: A classification tree analysis. *Children and Youth Services Review*, *30*, 1318-1327. DOI: 10.1016/j.childyouth.2008.03.014

[24]Snowden JA, Leon SC, Bryant FB, Lyons JS (2007). Evaluating psychiatric hospital admission decisions for children in foster care: An optimal classification tree analysis. *Journal of Child and Adolescent Psychology*, *36*, 8-18. DOI: 10.1080/15374410709336564

[25]Soltysik RC, Yarnold PR (2014). Hierarchically optimal classification tree analysis of adverse drug reactions secondary to warfarin therapy. *Optimal Data Analysis*, *3*, 23-24. URL: http://odajournal.com/2014/04/01/hierarchically-optimal-classification-tree-analysis-of-adverse-drug-reactions-secondary-to-warfarin-therapy/

[26]Stalans LJ, Hacker R, Talbot ME (2010). Comparing nonviolent, other-violent, and domestic batterer sex offenders: Predictive accuracy of risk assessments on sexual recidivism. *Criminal Justice and Behavior*, *37*, 613-628. DOI: 10.1177/0093854810363794

[27]Stalans LJ, Seng M (2006). Identifying subgroups at high risk of dropping out of domestic batterer treatment: The buffering effects of a high school education. *International Journal of Offender Therapy and Comparative Criminology*, *10*, 1-19. DOI: 10.1177/0306624X06290204

[28]Stoner AM, Leon SC, Fuller AK (2013). Predictors of reduction in symptoms of depression for children and adolescents in foster care. *Journal of Child and Family Studies*, *22*, DOI 10.1007/s10826-013-9889-9

[29]Suzuki H, Bryant FB, Edwards JD (2010). Tracing prospective profiles of juvenile delinquency: An optimal classification tree analysis. *Optimal Data Analysis*, *1*, 125-143. URL: http://odajournal.com/2013/09/19/tracing-prospective-profiles-of-juvenile-delinquency-and-non-delinquency-an-optimal-classification-tree-analysis/

[30]Yarnold PR (2014). Triage algorithm for chest radiography for community-acquired pneumonia of Emergency Department patients: Missing data cripples research. *Optimal Data Analysis*, *3*, 102-106. URL: http://odajournal.com/2014/09/02/triage-algorithm-for-chest-radiography-for-community-acquired-pneumonia-of-emergency-department-patients-missing-data-cripples-research/

[31]Yarnold PR, Soltysik RC, Collinge WC (2013). Modeling individual reactivity in serial designs: An example involving changes in weather and physical symptoms in fibromyalgia. *Optimal Data Analysis*, *2*, 37-42. URL: http://odajournal.com/2013/09/20/modeling-individual-reactivity-in-serial-designs-changes-in-weather-and-physical-symptoms-in-fibromyalgia/

2155-0182/10/$3.00

[32]Yarnold PR, Bryant FB, & Smith JH. (2013). Manual vs. automated CTA: Predicting freshman attrition. *Optimal Data Analysis*, *2*, 48-53. URL: http://odajournal.com/2013/09/20/manual-vs-automated-cta-predicting-freshman-attrition/

[33]Yarnold PR, Soltysik RC (2010). Manual vs. automated CTA: Optimal preadmission staging for inpatient mortality from *Pneumocystis cariini* pneumonia. *Optimal Data Analysis*, *1*, 50-54. URL: http://odajournal.com/2013/09/19/manual-vs-automated-cta-optimal-preadmission-staging-for-inpatient-mortality-from-pneumocystis-cariini-pneumonia/

[34]Coakley RM, Holmbeck GN, Bryant FB, Yarnold PR (2010). Manual vs. automated CTA: Predicting adolescent psychosocial adaptation. *Optimal Data Analysis*, *1*, 55-58. URL: http://odajournal.com/2013/09/19/manual-vs-automated-cta-psychosocial-adaptation-in-young-adolescents/

[35]Yarnold PR (2014). Increasing the likelihood of an ambivalent patient recommending the Emergency Department to others, *Optimal Data Analysis*, *3*, 89-91. URL: http://odajournal.com/2014/08/27/increasing-the-likelihood-of-an-ambivalent-patient-recommending-the-emergency-department-to-others/

[36] Bryant FB, Harrison PR (2013). How to create an ASCII inout data file for UniODA and CTA software. *Optimal Data Analysis*, *3*, 2-6. URL: http://odajournal.com/2013/09/20/how-to-create-an-ascii-input-data-file-for-unioda-and-cta-software/

[37]Yarnold, P.R. (2014). Illustrating how 95% confidence intervals indicate model redundancy. *Optimal Data Analysis*, *3*, 96-97. URL: http://odajournal.com/2014/08/31/illustrating-how-95-confidence-intervals-indicate-model-redundancy/

Author Notes

This study involved secondary data analysis of published de-identified data and was exempt from Institutional Review Board review.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646
USA