Addressing the Needs of Outlier Substance Abuse Outpatients

Madeleine Gleave, Sook Yee Leung, and Minal Singhee

Large Scale Data Analysis for Public Policy

December 12, 2015

Introduction

Across the healthcare industry, including substance abuse treatment programs, a disproportionate share of costs are driven by a few outlier patients, particularly for patients with chronic conditions (O'Connor, 2015). According to the National Institute of Drug Abuse, abuse of marijuana and illegal drugs has been on increasing since 2007 ("Nationwide Trends," 2015). Given that there is a growing number of substance abusers, more resources would be required of substance abuse clinics. According to French, Popovici, and Tapsell (2008), the top five costliest substance abuse treatments include "therapeutic communi[ties]," "methodone maintenance," "adult residential," "intensive outpatient," and "drug court" treatments (p.13, *Figure 1*). Therefore, identifying and examining the characteristics of substance abuse patients requiring longer treatment stays than other patients could be useful for treatment clinics and policymakers seeking to reduce costs and improve services.

Proposed Weekly and Episode Cost Bands

Modality	Weekly Cost	Episode Cost
Methadone maintenance	\$87 – \$112	\$4,277 – \$13,395
Non-methadone outpatient	\$74 – \$221	\$1,132 – \$2,099
Intensive (i.e., day treatment) outpatient	\$243 – \$598	\$1,384 – \$5,780
Adolescent outpatient	\$139 – \$281	\$1,517 – \$3,237
Drug court	\$34 – \$146	\$2,486 – \$4,888
Adult residential	\$607 – \$918	\$2,907 – \$11,260
Therapeutic community	\$569 – \$708	\$14,818 – \$32,361
In-prison therapeutic community	\$55 – \$71	\$1,249 – \$2,112

Notes: The cost bands correspond to the interquartile range for each modality (25th to 75th percentile). Estimates are not reported for "screening and brief intervention" and "adolescent residential" due to very small sample sizes for these modalities.

Figure 1. Cost estimates based on specific substance abuse treatments (French, Popovici, and Tapsell, 2008, p.13).

Given that the aforementioned modalities imply differing kinds of alternative treatment options, this project will focus on examining the characteristics that contribute to predicting outlier intensive outpatients. Patients in intensive outpatient treatments undergo at least two hours of treatment within a facility at least three days per week, but are not residentially housed at the facility (TEDS-D 2011 Codebook, 2011, p.18). Thus, since intensive outpatients are still living within their communities, insights from the characteristics of these patients in the dataset may provide insights for alternative or enhanced services based in the community. Using classification, attribute-relationship modeling, and clustering algorithms, as well as experimental anomaly detection methods, we explore how length of stay and, subsequently, treatment costs may be reduced, as well as how these outlier patients' needs can be addressed more efficiently by facility and/or complementary programs.

The Data

The Treatment Episode Data Set -- Discharges (TEDS-D) for the years 2006 to 2011 were found ("Treatment Episode Data Set - Discharges (TEDS-D) Series", 2015). Each dataset contains over one million records. Due to computational limitations, we chose to focus on the most recent dataset: TEDS, 2011 (ICPSR 35074) from the *Substance Abuse and Mental Health Data Archive* (Treatment Episode Data Set -- Discharges (TEDS-D), 2011). The dataset has 1,732,741 records on discharged patients from substance abuse treatment clinics that provide information to the "individual state administrative data systems" (TEDS-D 2011 Codebook, 2011, p.1). The dataset includes information on "treatment completion, length of stay in treatment, and demographic and substance abuse" for each reported patient, which results to a total of 65 variables (TEDS-D 2011 Codebook, 2011, p.1). The complete list is shown in *Figure 1*. Each record represents a unique patient in a substance abuse clinic across the United States.

ATTRIBUTE	DESCRIPTION
CASEID	CASE IDENTIFICATION NUMBER
DISYR	YEAR OF DISCHARGE
AGE	AGE (RECODED)
GENDER	SEX
RACE	RACE
ETHNIC	ETHNICITY (HISPANIC ORIGIN)
MARSTAT	MARITAL STATUS
EDUC	EDUCATION
EMPLOY	EMPLOYMENT STATUS
DETNLF	DETAILED 'NOT IN LABOR FORCE' CATEGORY
PREG	PREGNANT AT TIME OF ADMISSION
VET	VETERAN STATUS
LIVARAG	LIVING ARRANGEMENTS
PRIMINC	SOURCE OF INCOME/SUPPORT
ARRESTS	NUMBER OF ARRESTS IN 30 DAYS PRIOR TO ADMISSION
STFIPS	CENSUS STATE FIPS CODE
CBSA	FIPS 2000 CBSA CODE
PMSA	FIPS 1990 MSA CODE
REGION	CENSUS REGION
DIVISION	CENSUS DIVISION
SERVSETD	SERVICE SETTING AT DISCHARGE
METHUSE	MEDICATION-ASSISTED OPIOID THERAPY
DAYWAIT	DAYS WAITING TO ENTER TREATMENT
REASON	REASON FOR DISCHARGE
LOS	LENGTH OF STAY
PSOURCE	PRINCIPAL SOURCE OF REFERRAL
DETCRIM	DETAILED CRIMINAL JUSTICE REFERRAL

ATTRIBUTE	DESCRIPTION
CASEID	CASE IDENTIFICATION NUMBER
DISYR	YEAR OF DISCHARGE
AGE	AGE (RECODED)
GENDER	SEX
RACE	RACE
ETHNIC	ETHNICITY (HISPANIC ORIGIN)
MARSTAT	MARITAL STATUS
EDUC	EDUCATION
EMPLOY	EMPLOYMENT STATUS
DETNLF	DETAILED 'NOT IN LABOR FORCE' CATEGORY
PREG	PREGNANT AT TIME OF ADMISSION
VET	VETERAN STATUS
LIVARAG	LIVING ARRANGEMENTS
PRIMINC	SOURCE OF INCOME/SUPPORT
ARRESTS	NUMBER OF ARRESTS IN 30 DAYS PRIOR TO ADMISSION
STFIPS	CENSUS STATE FIPS CODE
CBSA	FIPS 2000 CBSA CODE
PMSA	FIPS 1990 MSA CODE
REGION	CENSUS REGION
DIVISION	CENSUS DIVISION
SERVSETD	SERVICE SETTING AT DISCHARGE
METHUSE	MEDICATION-ASSISTED OPIOID THERAPY
DAYWAIT	DAYS WAITING TO ENTER TREATMENT
REASON	REASON FOR DISCHARGE
LOS	LENGTH OF STAY
PSOURCE	PRINCIPAL SOURCE OF REFERRAL
DETCRIM	DETAILED CRIMINAL JUSTICE REFERRAL

ATTRIBUTE	DESCRIPTION
TRNQFLG	OTHER NON-BENZODIAZEPINE TRANQUILIZERS REPORTED AT ADM
BARBFLG	BARBITURATES REPORTED AT ADM
SEDHPFLG	OTHER NON-BARBITURATE SEDATIVES/HYPNOTICS REPORTED AT ADM
INHFLG	INHALANTS REPORTED AT ADM
OTCFLG	OVER-THE-COUNTER MEDICATION REPORTED AT ADM
OTHERFLG	OTHER DRUG REPORTED AT ADM
ALCDRUG	SUBSTANCE ABUSE TYPE
DSMCRIT	DSM DIAGNOSIS
PSYPROB	PSYCHIATRIC PROBLEM IN ADDITION TO ALCOHOL/DRUG PROBLEM
HLTHINS	HEALTH INSURANCE
PRIMPAY	EXPECTED/ACTUAL PRIMARY SOURCE OF PAYMENT

Figure 1. The entire list of TEDS-D, 2011 variables (TEDS-D 2011 Codebook, 2011, *Appendix A*)

Patients in this dataset received different modes of treatment. In accordance with the aforementioned focus on intensive outpatient treatment patients, we chose to focus on patients who were "ambulatory, intensive outpatient[s]" as shown in *Figure 2* (TEDS-D 2011, n.d., p.19). In the dataset, this group of patients were SERVSETD= 6.

Value	Label	Unweighted Frequency	%
1	DETOX, 24 HR, HOSPITAL INPATIENT	62579	3.6 %
2	DETOX, 24 HR, FREE-STANDING RESIDENTIAL	273616	15.8 %
3	REHAB/RES, HOSPITAL (NON-DETOX)	6399	0.4 %
4	REHAB/RES, SHORT TERM (30 DAYS OR FEWER)	169655	9.8 %
5	REHAB/RES, LONG TERM (MORE THAN 30 DAYS)	132171	7.6 %
6	AMBULATORY, INTENSIVE OUTPATIENT	218013	12.6 %
7	AMBULATORY, NON-INTENSIVE OUTPATIENT.	851629	49.1 %
8	AMBULATORY, DETOXIFICATION	18312	1.1 %
	Missing Data		
-9	MISSING/UNKNOWN/NOT COLLECTED/INVALID	367	0.0 %
	Total	**1,732,741**	**100%**

Based upon 1,732,374 valid cases out of 1,732,741 total cases.

Figure 2. Treatment Type Groups Summary and Label Description (SERVSETD) [TEDS-D 2011 Codebook, 2011, p.19].

The variable of interest was the length of stay (LOS) as shown in *Figure 3*. This variable provides the total number of days the patient stayed in the clinic. The length of stay variable was originally partially numeric, providing the exact number of days from 1 to 30, and partially categorical, providing ranges of stay duration beyond 30 days.

LOS: LENGTH OF STAY

Describes the length of the treatment episode (in days). Length of stay was computed using the date of admission and the date of last contact. One day is added to all outpatient discharges, so that the first day and last day of outpatient treatment are counted.

- 1 to 30: Data values in the 1-30 range represent the actual computed number of days the client spent in treatment

- 31: Length of stay is between 31 and 45 days.

- 32: Length of stay is between 46 and 60 days.

- 33: Length of stay is between 61 and 90 days.

- 34: Length of stay is between 91 and 120 days.

- 35: Length of stay is between 121 and 180 days.

- 36: Length of stay is between 181 and 365 days.

- 37: Length of stay is greater than 365 days.

Figure 3. Description of the original length of stay variable (TEDS-D 2011 Codebook, 2011, p. 23).

Data Preparation

Due to computational limitations, the entire dataset could not be used to train a model. *Weka* was used to create algorithmic models to examine the data. Before conducting analysis on intensive outpatients within the TEDS-D 2011 dataset, 30% (519,823 records) of the data was set aside as unseen data that was not used to train an algorithmic model. Seventy-percent of the data (1,212,918 instances) was allotted to train an algorithmic model. Nevertheless, limited computational power allowed only 15,000 records to be randomly chosen and designated as the model training set and 1,000 records to be randomly chosen from the unseen data set to test the final algorithmic model.

In accordance with the research focus of this project, the 15,000 records chosen from the model training data set received the intensive outpatient treatment (SERVSETD = 6). Furthermore, given that the focus of this project is to classify and characterize patients with unusually long stays, the LOS variable was recoded to create an anomalous group with treatment durations beyond 180 days (longer) and a normal group with treatment durations of and below 180 days (shorter). *Figure 4* depicts the number of the newly recoded LOS variable groups, which contain 88.06% shorter (or normal length) stays and 11.94% longer (or anomalous length) stays.

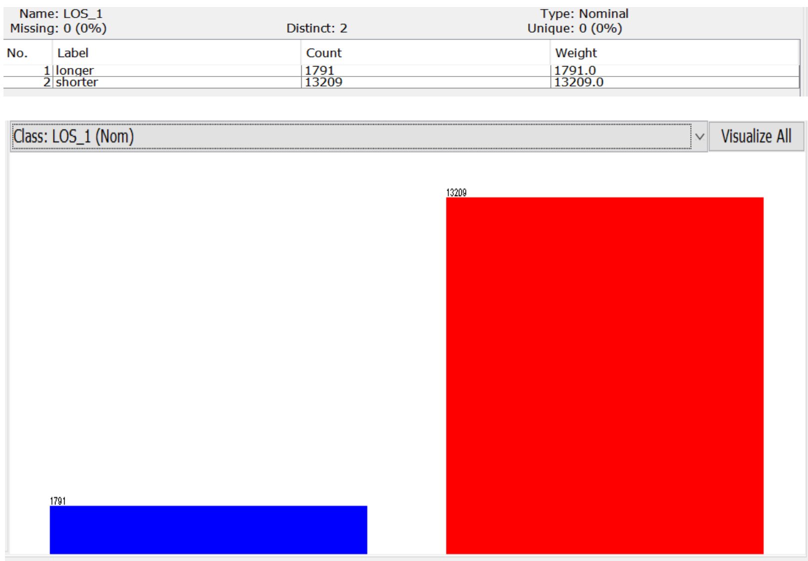

Figure 4. Recoded nominal LOS.

A smaller subset of features were selected from the original dataset to reduce noisy features that can reduce the accuracy of the trained model and to create a smaller sized dataset that may be more computationally manageable via *Weka*. Before employing feature selection, high arity attributes, such as CASEID and some location attributes (e.g., CBSA-FIPS 2000 CBSA CODE and PMSA- FIPS 1990 MSA CODE), were removed. The feature selection method used was information gain. The Weka setting used was InfoGainAttributeEval with the Ranker search method via 10-fold cross validation (*Figure 5*).

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.016 +- 0.001      1   +- 0       10 DIVISION
 0.009 +- 0          2   +- 0        7 PRIMINC
 0.006 +- 0          3   +- 0       13 REASON
 0.006 +- 0          4   +- 0       14 PSOURCE
 0.004 +- 0          5.3 +- 0.46     5 DETNLF
 0.004 +- 0          6.6 +- 1.02     3 MARSTAT
 0.004 +- 0          6.7 +- 1.27    15 SUB1
 0.003 +- 0          9.1 +- 1.37    25 HLTHINS
 0.003 +- 0          9.4 +- 1.8     26 PRIMPAY
 0.003 +- 0          9.9 +- 2.07    19 SUB3
 0.003 +- 0         10.9 +- 1.87    24 DSMCRIT
 0.003 +- 0         12.4 +- 1.69    12 DAYWAIT
 0.002 +- 0         12.9 +- 1.87    11 METHUSE
 0.003 +- 0         12.9 +- 1.45     2 RACE
 0.002 +- 0         14.1 +- 1.45    18 FREQ2
 0.002 +- 0         16.3 +- 0.64     9 REGION
 0.002 +- 0         16.6 +- 0.8      4 EMPLOY
 0.001 +- 0         18.8 +- 0.98    16 ROUTE1
 0.001 +- 0         19.5 +- 1.28    17 FREQ1
 0.001 +- 0         19.6 +- 1.02     6 LIVARAG
 0.001 +- 0         20.2 +- 1.33    23 OTHERFLG
 0.001 +- 0         22.8 +- 0.75    22 STIMFLG
 0.001 +- 0         23.4 +- 1.28    21 MTHAMFLG
 0.001 +- 0         23.5 +- 1.28    20 HERFLG
 0.001 +- 0         24.1 +- 0.94     8 ARRESTS
 0     +- 0         26   +- 0        1 GENDER
```

Figure 5. Feature selection output

As a shown above, a total of 26 attributes were presented in *Weka* and all 26 features were kept for creating the final model. Despite the low average merit, GENDER was kept in the dataset because this attribute is of domain-based interest: the relationship between gender and substance abuse in the context of the other attributes may yield interesting insights.

Model Selection

Baseline

Before selecting the best classification model for the training data, a baseline model was assessed. The ZeroR model, which predicted the majority class (i.e., "shorter"), was used. For the current research question the ZeroR model represented the least ideal classification model, since the model would miss the predicting the anomalous class label (i.e., "longer"), which is not a majority. The ZeroR is a rather simple model since no other attributes, except the variable of interest is used to create the model. While the model accuracy (88.06%) is high, since the "shorter" class comprises of the majority of the dataset, the accuracy metric is not useful for detecting the "longer" class, which is the minority and the anomaly within the dataset. Upon closer examination, the ROC area and recall were found to be more useful metrics in assessing how well a model detected the anomalous class. The ROC area indicated how well a classifier distinguished a particular LOS label from the other LOS label ("The area under the ROC curve, n.d.). An ROC area of 0.5 indicates that the model is predicting the particular class label randomly and an ROC area of 1 indicates that the model is predicting the particular class label perfectly considering the presence of the other class labels. The recall for a particular LOS label would indicate the percentage of the LOS label being predicted by the model out of all of the actual instances of the LOS label in the training dataset or the true positives over the true positives and the false negatives in the training data ("Recall and precision in classification", 2013).

As shown in *Figure 6*, the ZeroR model does not perform well in predicting the LOS = "longer" class label. When only considering the recall for LOS = "longer", the recall is 0, which means that the ZeroR model is not predicting any true LOS= "longer" labels. Moreover, an ROC area of 0.5 indicates that the model randomly distinguishes the LOS="longer" class label from the other class label. The baseline model provides insight into how other algorithmic models may improve predictions.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13209               88.06   %
Incorrectly Classified Instances      1791               11.94   %
Kappa statistic                          0
Mean absolute error                      0.2103
Root mean squared error                  0.3243
Relative absolute error                100      %
Root relative squared error            100      %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances            15000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      longer
                 1         1          0.881     1         0.937      0.5      shorter
Weighted Avg.    0.881     0.881      0.775     0.881     0.825      0.5

=== Confusion Matrix ===

     a     b   <-- classified as
     0  1791 |     a = longer
     0 13209 |     b = shorter
```

Figure 6. ZeroR output for the training dataset.

Model Selection

Different algorithmic models were trained with the training set data using 10-fold cross validation. Overall, the models trained with the reduced attributes (*Figure 8*) were more accurate than the models trained with all features except for CASEID (*Figure 7*). Similar to the ZeroR model, the decision tree (J48) model and the rule learning algorithm (JRip) learned a model that was heavily based on LOS to predict the majority class label (i.e., "shorter"). The aforementioned is reflected in low recall value, indicating a lack of predictions of actual "longer" class labels, and the approximately 0.5 AUC value, indicating that the models randomly distinguishes the LOS="longer" class label from the other class label.

(All features)	ZeroR	Naïve Bayes	IBk (Knn = 1)	BayesNet (Parents = 2)	JRip	Decision Tree (J48)
Accuracy	88.06%	85.76%	81.59%	85.27%	87.73%	88.06
AUC	0.5	0.687	0.594	**0.707**	0.5	0.5
Recall	0	0.124	0.258	**0.197**	0.056	0

Figure 7. Accuracy, ROC Area (AUC), and Recall metrics across trained models in predicting the "longer" class label for models trained on all features except for CASEID (attribute was unique to each row).

(26 features)	ZeroR	Naïve Bayes	IBk (Knn = 1)	BayesNet (Parents = 2)	JRip	Decision Tree (J48)
Accuracy	88.06%	86.86%	82.8%	85.76%	87.97%	88.06
AUC	0.5	0.695	0.635	**0.713**	0.525	0.5
Recall	0	0.095	0.26	**0.177**	0.056	0

Figure 8. Accuracy, ROC Area (AUC), and Recall metrics across trained models in predicting the "longer" class label on the reduced features training set.

Upon comparing the Naïve Bayes, K nearest neighbors (KNN = 1), and Bayesian Network (BayesNet: parents = 2) models, the Bayesian Network model was chosen based on the following three criteria: performance, potential extent of underfitting/overfitting, and interpretability. Regarding performance, the Bayesian network had the best ROC area performance of 0.713, which falls within the "fair" range of 0.7 to 0.8 as opposed to the other models which falls within the "poor" range of 0.6 to 0.7 (http://gim.unmc.edu/dxtests/roc3.htm). Nevertheless, the Bayesian Network had the second best recall performance. In terms of recall, the K nearest neighbor model had best recall, but the lower ROC area may indicate that the K nearest neighbor model does not distinguish between the two classification labels as well as the Bayesian network, suggesting that the K nearest neighbor model may be overfitting to the training data at the KNN = 1 setting. In terms of interpretability, the Bayesian Network model may be more informative than KNN model since the interrelated probabilities among the attributes can be shown graphically to describe characteristics affecting LOS. The K nearest neighbor model only provides record by record distances that do not explicitly note relationships among attributes. In other words, the Bayesian Network model provides more information about the characteristics of anomalous intensive outpatient patients that can be used towards further policy considerations.

Final test result

A final test of the trained model on unseen data was conducted by applying the model to 1,000 randomly selected unseen TEDS-D 2011 data instances. The aforementioned practice is similar to real-world situations of applying trained machine learning algorithmic models to unseen data to assess the applicability of the model to future data. Moreover, due to computational limitations a subset of the original data was randomly selected via R, this final test was conducted to ascertain whether the randomly selected instances used to train the Bayes Net model would perform similarly on another subset of the TEDS-D 2011 data. According to *Figure 9*, the trained model's test set performance yield an ROC area of 0.723 and recall of 0.192 when predicting the anomalous "longer" class label, which is similar to the trained model's performance on the training data, which resulted to an ROC area of 0.713 and recall of 0.177 when predicting the anomalous "longer" class label in *Figure 8*. Similar performance on the unseen data indicates that the Bayes Net model was trained with records that were representative of the TEDS-D 2011 data.

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     trmt6test1k-weka.filters.unsupervised.attribute.Remove-R1-4,19-20-weka.filters.uns
Instances:     unknown (yet). Reading incrementally
Attributes:   27

=== Summary ===

Correctly Classified Instances         857               85.7    %
Incorrectly Classified Instances       143               14.3    %
Kappa statistic                          0.1707
Mean absolute error                      0.1988
Root mean squared error                  0.327
Coverage of cases (0.95 level)          98.4    %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.192     0.052     0.333     0.192     0.243      0.723    longer
                 0.948     0.808     0.896     0.948     0.921      0.723    shorter
Weighted Avg.    0.857     0.718     0.828     0.857     0.84       0.723

=== Confusion Matrix ===

   a   b   <-- classified as
  23  97 |   a = longer
  46 834 |   b = shorter
```

Figure 9. Applying the trained Bayes Net (maxNumofParents = 2) model to unseen data of 1,000 instances.

Analysis

Bayesian Network

Because we are already able to identify the characteristics that qualify the individuals in our dataset as outliers, i.e. they have a significantly longer length of stay, the most valuable element of our model is determining what the patterns and relationships between other characteristics are for those individuals. A Bayesian Network (Bayes Net) model is particularly well-suited for this. The Bayes Net encodes relationships between attributes through conditional probabilities, helping us to see causal links and sequencing in those attributes. These probabilities also help us determine which subsets of related characteristics have higher or lower probabilities, and the degree to which those probabilities depend on interactions with "parent" nodes and attributes.

Our selected Bayes Net model, as described above, maps the 26 features with the highest information gain. As seen in *Figure 10*, there are two main streams of relationships: demographic attributes on the left, and regionally-driven attributes on the right, many of which are more clinical in nature. Of course, there are many intersecting lines in between these nodes, and most have multiple parents. All share the Length of Stay (LOS) classifying node as a parent, suggesting the intermediary relationships modeled cannot fully explain the predictive probabilities of each attribute on length of stay. Nonetheless, the model reveals some interesting and useful dependencies in understanding the factors that determine a patient's likelihood to have an extended stay in treatment.

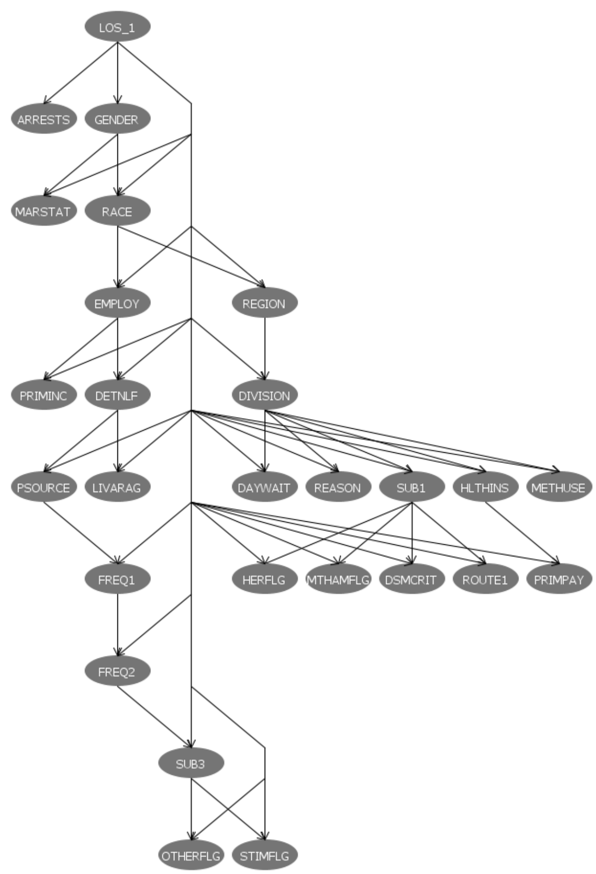

Figure 10. Visualization of Bayesian Network with Reduced Features

First, we can see that the influence of demographic characteristics cascades from a patient's gender. Looking at the posterior probabilities at this node, we observe that women have higher probability of longer stays (.446) than shorter stays (.42) and that the probabilities are reversed for men. However, given that gender had essentially zero information gain in our feature selection (but was included due to its substantive value as a key demographic characteristic) these differences are not likely to be significant. However, a child node of gender is race, and the interaction between these shows other interesting patterns. Although males overall are less likely to have longer stays, black males are .05 more likely to have longer rather than shorter stays in treatment. Black women, likewise, are .03 more likely to have longer stays.

The next set of nodes stemming from the demographic line are related to employment and living arrangements. The highest posterior probabilities for longer stay for the EMPLOY node are for those patients who are unemployed and not in labor force; unsurprisingly, the detailed reason not in labor force (DETNLF) is closely linked to the “Not in Labor Force” category of the parent node, EMPLOY. Furthermore, a patient’s living arrangements is a child of DETNLF, demonstrating the importance of employment status on other social conditions. Considering outpatient individuals are by nature not living in the treatment center, their living arrangements are particularly important aspects of their overall treatment episode conditions. This sequence of related attributes highlights one of the highest differential probabilities for longest stay, which is for patients in dependent living arrangements and not working. Specifically, dependent-living homemakers have .115 higher probability of longer length of stay in treatment. This could indicate a population of stay-at-home parents, likely the primary caregivers, who may be choosing an extended stay in outpatient treatment rather than a more intensive, short-term residential treatment that would require them to be away from the home. Additionally, retired or disabled patients living in dependent situations are .07 more likely to have longer stays in treatment. This could capture a deficiency of external social services that make it harder for these patients to complete treatment in a shorter amount of time.

On the other major line in the Bayes Net, we observe that geographic attributes have a strong influence over many other characteristics, including types of substance abuse. In our feature selection, these geographic attributes also had some of the highest information gain, even when they captured the hierarchical groupings of the same measure (i.e., regions are groupings of census divisions which are a grouping of states). While the full reasons for these relationships are not clear at this stage of our research, some of the parent/child interactions show that differences in length of stay may vary based on regional policies and trends. For example, the status/type of a patient’s health insurance (HLTHINS) is a child of DIVISION, suggesting different areas may have different health insurance requirements, companies, and coverage for outpatient substance abuse treatments. It is also true that the prevalence of certain addictive substances varies by region—for example, the recent epidemic of heroin use in New England communities. Thus, the interaction between geography and substance-related attributes is helpful to map.

Clustering

To further explore the similarities between patients with longer treatment stays, we employed a k-Means clustering methodology. Unfortunately due to the high dimensionality of our data and the low prevalence of the longer-stay class, we had to include a very high number of cluster points to see any with predominantly longer-stay patients, with k=31 being the minimum threshold. *Figure 11* shows the attribute modes for the one cluster identified at this level. As we can see, it appears to center around male, white meth addicts who are being referred by the court system. However, we cannot draw too strong of conclusions from our clustering methods, because even this cluster only has 59% longer-staying patients.

Attribute	Mode
GENDER	Male
RACE	White
MARSTAT	Divorced/widowed
EMPLOY	Full time
DETNLF	(missing, b/c employed)
LIVARAG	Independent living
PRIMINC	Wages/Salary
ARRESTS	None
REGION	West
DIVISION	Mountain
METHUSE	No
DAYWAIT	Less than one
REASON	Completed treatment
PSOURCE	Court / Criminal Justice Referral / DUI/DWI
SUB1	Meth
ROUTE1	Injection
FREQ1	None in past month
FREQ2	None in past month
SUB3	None
HERFLG	No
MTHAMFLG	Yes
STIMFLG	No
OTHERFLG	No
DSMCRIT	Other substance dependence
HLTHINS	None
PRIMPAY	Other government payments (likely state split payments)

Figure 11. Modes of first-appearing longer-stay cluster, at k=31.

Anomalous Subgroup Scan Method

Based on the Bayesian Network model, the probabilities for predicting the actual LOS label ("longer" or "shorter") was extracted for each of the 15,000 training instances via *Weka's* prediction output in conjunction with the accompanying attributes from the reduced feature set. Via an anomalous subgroup scan method, two most anomalous subgroups were found: (1) a group makes the most biased probability estimates towards predicting "shorter" and (2) a group that has the high number of prediction errors than statistically expected [Zhang, email communication, December 2015- Appendix A]. While the first subgroup consisted of 1.5% (223 instances) of the training dataset, insights on the attributes within this subgroup indicate characteristics that may indicate characteristic based on which attributes "shorter" patient cases were over-predicted. This subgroup over-predicted "shorter" for 92% of patient cases when in actuality only 72% of the cases where labelled "shorter" [Zhang, email communication, December 2015-- Appendix A].

As shown in figure 10, three racial groups (i.e., African Americans, White, and Hawaiian and other Pacific Islanders), three employment statuses (i.e. unemployed, not in workforce, full-time), two arrest numbers (i.e., no arrests, two arrests), and treatment characteristics (i.e., not on medication assisted opioid therapy) were in this subgroup that over-predicted of "shorter" patient stays. Cases with the aforementioned characteristics reflect attributes used to predict shorter patient stays, but given that not all patients with the attributes are labelled short (i.e., completely red), such as RACE = white (5) or EMPLOY = full time (1), overreliance on using such attribute labels to make predictions results to incorrect over-prediction of shorter. The subgroups were examined in *R* and then extracted into *Weka* for further analysis as shown in *Figure 12*.

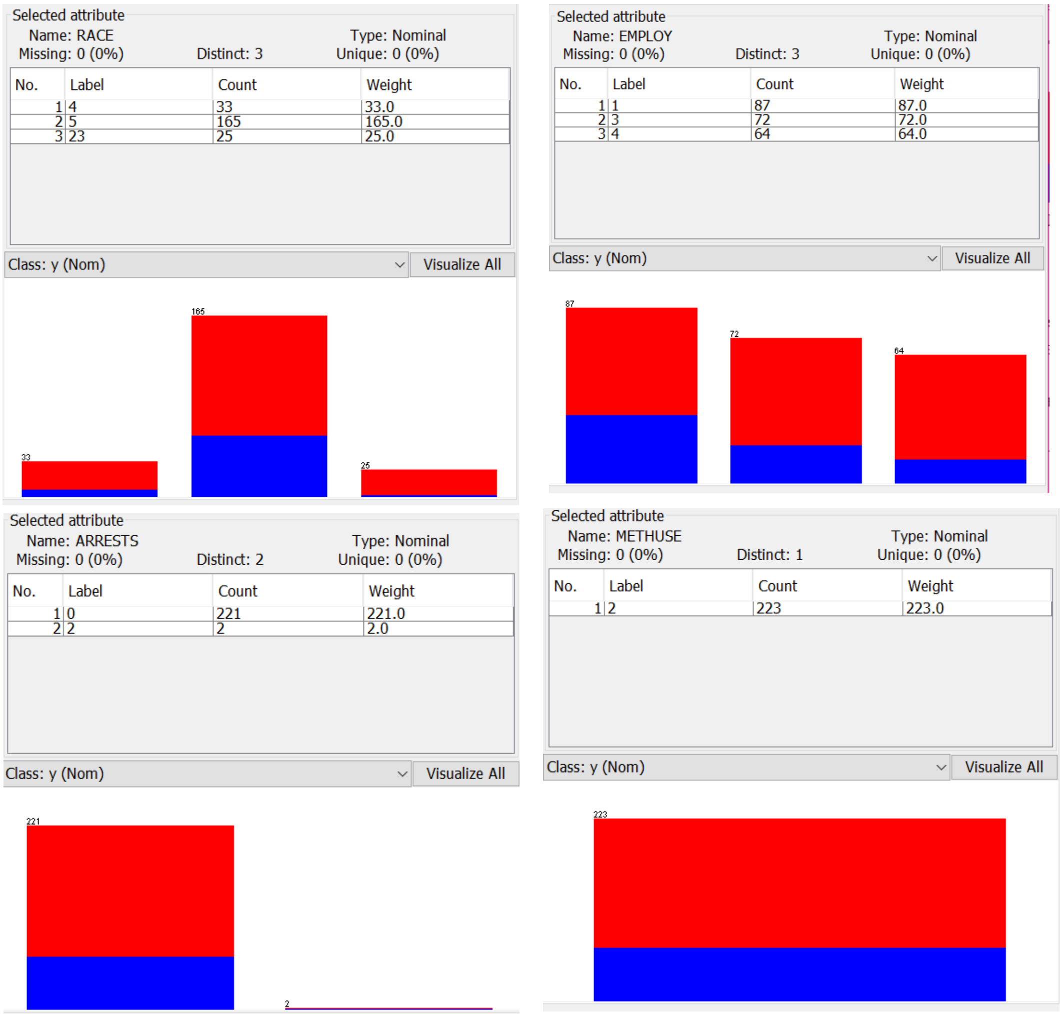

Figure 12. [Upper Left] Three racial groups were over-predicted to have shorter stays (mainly red). Blue indicates longer stays. 4 = Black or African American, 5 = White, 23 = Native Hawaiian or other Pacific Islander (TEDS-D 2011 Codebook, 2011, p.4). [Upper Right] Three employment categories were over-predicted to have shorter stays (mainly red). Blue indicates longer stays. 1 = full time, 3 = unemployed, 4 = not in workforce (TEDS-D 2011 Codebook, 2011, p.7). [Lower left] The number of arrests prior to clinic admission. (TEDS-D 2011 Codebook, 2011, p.51). [Lower right] Medication-assisted opioid therapy. METHUSE = 2 means that no medication assisted opioid therapy was used. (TEDS-D 2011 Codebook, 2011, p.51).

While the second subgroup consisted of 1% (120 instances) of the training dataset, this group had an error rate of 37%, which is statistically significant compared to other subgroups of the training data [Zhang, email communication, December 2015]. For instance, the previous subgroup had an error rate of 29%. The overall dataset contains 12% prediction errors [Zhang, email communication, December 2015]. Again, the “shorter” is predicted more than “longer” so patterns within this group indicates based on which attributes “longer” patient cases were misclassified. Similar to the groups noted in figure 10, two racial groups (i.e., African Americans and White), three employment statuses (i.e. unemployed, not in workforce, full-time), one arrest number (i.e., no arrests), and treatment characteristics (i.e., not on medication assisted opioid therapy) were in this subgroup that over-predicted of "shorter" patient stays. In *Figure 13*, cases with the aforementioned characteristics reflect attributes used to predict shorter patient stays, but given that not all patients with the attributes are actually labelled short (i.e., completely red), such as RACE = white (5) or EMPLOY = full time (1), overreliance on using such attribute labels to predict shorter (the majority predicted class of this subgroup) result to incorrect over-prediction of shorter.

Figure 13. [Upper Left] Three racial groups were over-predicted to have shorter stays (mainly red). Blue indicates longer stays. 4 = Black or African American and 5 = White(TEDS-D 2011, n.d., p.4). [Upper Right] Three employment categories were over-predicted to have shorter stays (mainly red). Blue indicates longer stays. 1 = full time, 3 = unemployed, 4 = not in workforce (TEDS-D 2011, n.d., p.7). [Lower left] The number of arrests prior to clinic admission. (TEDS-D 2011, n.d., p.51). [Lower right] Medication-assisted opioid therapy. METHUSE = 2 means that no medication assisted opioid therapy was used. (TEDS-D 2011, n.d., p.51).

In accordance with the Bayesian Network analysis, the aforementioned attributes and accompanying labels in *Figures 12* and *13* indicate attribute labels that may be indicative of longer stays. The cases sharing the characteristics that are not decidedly indicative of "shorter" stays (i.e., has a completely red bar) might be misclassified as contributors to shorter stays since the training dataset does not have an equal amount of information of longer patient stay records in comparison to shorter stay records.

Conclusion

Our research is an attempt to understand not only the predicting attributes of patients with outlier lengths of stay in substance abuse treatment, but also the relationships between these attributes. A Bayesian Network model provided the best predictive power and allowed us to map those causal relationships. Based on the results in our framework, we make the following recommendations to policymakers engaged in the substance abuse treatment field.

First, treatment centers and community organizations should work together to focus on demographic subset populations with specific vulnerabilities. As laid out in the framework of the Treatment Improvement Protocols for Intensive Outpatient Treatment published by the Center for Treatment of Substance Abuse, there are many enhanced services beyond the core clinical programs and pharmacological solutions that treatment centers can offer to facilitate faster, more effective, and more relevant treatment engagements for their patients (CTSA, 2006). Treatment centers should be particularly aware of patients who are living in dependent situations as the primary care-giver, and provide more comprehensive services that would ease and aid their situation. This is particularly important given the significant impact cases of substance-abusing parents can have on child welfare (Grella et al., 2006). These services could include child care and parenting education support. Additionally, treatment centers should coordinate with welfare and other social services to provide for needs of elderly or disabled and dependent patients, such as offering accessible transportation services.

Second, the broader substance abuse treatment policy community should investigate further into differences between state and regional policies regarding treatment requirements and outpatient services. Where possible, these policies should be aligned with regional needs and demographic populations. On both of these fronts, the effort of policymakers and treatment facility management should be dual in shortening outpatients' length of stay: both to reduce costs for the facility, making a more efficient system which can serve more patients, and to improve service delivery for patients, making their treatment experience more effective.

Further research is needed to understand the complexity of these factors. With such a rich dataset, there are many possibilities for understanding a patient's experience and methods for improvement treatment. One area of further research would be to examine the factors that influence successful completion of treatment, not treatment episode length in general. It is possible that those with longer treatment stays are doing so at the recommendation of medical professionals, whereas there is another subset of individuals who drop out early. Understanding the differences and predictive factors between these two groups would yield important additional context for policymakers and management when considering options to improve substance abuse treatment.

References

Center for Substance Abuse Treatment. Substance Abuse: Clinical Issues in Intensive Outpatient Treatment. Rockville (MD): Substance Abuse and Mental Health Services Administration (US); 2006. (Treatment Improvement Protocol (TIP) Series, No. 47.) Chapter 4. Services in Intensive Outpatient Treatment Programs. Retrieved December 12, 2015 from: http://www.ncbi.nlm.nih.gov/books/NBK64094/

French, M.T., Popovici, I., Tapsell, L. *J Subst Abuse Treat.* 2008 December ; 35(4): 462–469.

Grella, Christine E., Yih-Ing Hser, Yu-Chuang Huang, Mothers in substance abuse treatment: Differences in characteristics based on involvement with child welfare services, Child Abuse & Neglect, Volume 30, Issue 1, January 2006, Pages 55-73

O'Connor, M. (2015, August 30). What will it take to shift the 80-20 rule for healthcare? - MedCity News. Retrieved December 12, 2015, from http://medcitynews.com/2015/08/80-20-rule-for-healthcare/

Nationwide Trends. (2015, June 1). Retrieved December 12, 2015, from http://www.drugabuse.gov/publications/drugfacts/nationwide-trends

Recall and precision in classification. (2013, June 26). Retrieved December 12, 2015, from http://stats.stackexchange.com/questions/62621/recall-and-precision-in-classification

The Area Under an ROC Curve. (n.d.). Retrieved December 12, 2015, from http://gim.unmc.edu/dxtests/roc3.htm

TEDS-D 2011 Codebook. (2011). Retrieved December 12, 2015, from http://www.icpsr.umich.edu/cgi-bin/file?comp=none&study=35074&ds=1&file_id=1153830&path=ICPSR

Treatment Episode Data Set - Discharges (TEDS-D) Series. (2015). Retrieved December 12, 2015, from http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/238

Treatment Episode Data Set -- Discharges (TEDS-D), 2011. *United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality.* ICPSR35074-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2014-05-12. http://doi.org/10.3886/ICPSR35074.v1

Appendix A

Hi,

Apologies for the delay on this. I got caught up and also tried something that took way too long at first. It was much faster when I just searched for the most anomalous subpopulation.

In the Rdata file I attach, I find two anomalous subpopulations. One, is a subpopulation that has the most biased probability estimates. This group has a high estimated probability of short stay (y = 1) (avg 92%), but actually is only observed to have 72% short stay.
Two, is a subpopulation that has the highest number of prediction errors than we'd statistically expect. It has 37% of prediction errors while the overall dataset has 12% of data are prediction errors.

Following is some R code to display some of these results.

A quick primer on this subgroup. In the code below, you'll see early on what defines these two subgroups. What follows are values for each of the ~25 attributes that were selected to define the subgroup.

How to interpret this? It's a little tricky. The best way would be to look at each attribute and see what is restrictive. For example, if for some variable, like Age, the subgroup only chose a small set of values (e.g. choosing only people < age 15), that means that variable is restrictive. The opposite example, a non-restrictive set of values, would be all ages > 18.

This might be tricky for you to get done before your presentation (sorry I couldn't get to my computer earlier tonight). However, maybe something will jump out to you, and you can mention it or include it in your report.
These two groups are small but not tiny — 1.5% and 1% of the overall data respectively.

Definitely let me know what you think. It'd be easy to get a Top 10 or Top 15 list of subpopulations, right now, I only got a Top 1 though.
Thanks,
Zhe

Here's the R code to process the Rdata file:

```
# Load in Data
load('substanceScanResults.Rdata')

# What defines the over-estimated probability subgroup
underFull$bestLocalConditions

# What defines the high prediction error subgroup
errorFull$bestLocalConditions

# Whole Dataset
mean(dat$y)
mean(dat$pred)
```

```
# overall error rate (12%)
mean(abs(round(dat$pred) - dat$y))

# Subgroup that is over-estimated (more likely than estimated to have extended stay)
# percent of the total dataset (1.5% of dataset)
nrow(chosenUnderSubset) / nrow(dat)
# bias in estimated probability
mean(chosenUnderSubset$y)
mean(chosenUnderSubset$pred)
# subgroup error rate (29%)
mean(abs(round(chosenUnderSubset$pred) - chosenUnderSubset$y))


# Subgroup that has more errors than we would statistically expect
# percent of the total dataset (1% of dataset)
nrow(chosenErrorSubset) / nrow(dat)
# bias in subgroup
mean(chosenErrorSubset$y)
mean(chosenErrorSubset$pred)
# subgroup error rate (37%)
mean(abs(round(chosenErrorSubset$pred) - chosenErrorSubset$y))
```