Dravet Prediction Model

March 17, 2017

### Objective

Find algorithm/equation to best differentiate severe (Dravet and infantile spasm and Lennox Gastaut) from mild state (childhood absence epilepsy).

#### Steps

1. Data Management
2. Model Building
3. Evaluation of Model
4. Conclusion/Summary

#### Step 1: Data Management

***1a : Merge Chronic Comorbidity data with Patient Record Data***

***1b : Merge Medication data with Patient Record Data***

## 'data.frame': 191 obs. of 63 variables:  
## $ UNIQUE\_ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ freq.DiazePAM : num 1 1 2 0 3 0 0 13 2 2 ...  
## $ freq.CloBAZam : num 0 5 16 19 7 9 3 7 5 12 ...  
## $ freq.Divalproex Sodium : num 0 5 0 0 4 0 1 0 0 0 ...  
## $ freq.LevETIRAcetam : num 0 0 16 0 0 0 0 4 0 0 ...  
## $ freq.Midazolam HCl : num 0 0 1 4 2 1 0 0 1 0 ...  
## $ freq.Sodium Chloride : num 0 0 4 12 0 0 0 1 0 0 ...  
## $ freq.ClonazePAM : num 0 0 0 8 0 0 3 0 2 0 ...  
## $ COHORT : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ neuromusc\_ccc : int 0 1 1 1 1 1 1 1 1 1 ...  
## $ cvd\_ccc : int 0 1 0 0 0 0 1 1 0 0 ...  
## $ respiratory\_ccc : int 0 1 0 1 0 0 0 1 0 1 ...  
## $ renal\_ccc : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GI\_ccc : int 0 1 0 1 1 0 0 1 1 0 ...  
## $ hemato\_immu\_ccc : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ metabolic\_ccc : int 0 0 0 1 0 0 0 1 0 0 ...  
## $ congeni\_genetic\_ccc : int 0 1 1 1 1 0 1 0 0 1 ...  
## $ malignancy\_ccc : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ neonatal\_ccc : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ tech\_dep\_ccc : int 0 1 0 1 1 0 1 1 0 0 ...  
## $ transplant\_ccc : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ num\_ccc : int 0 5 2 5 3 1 3 5 3 3 ...  
## $ ccc\_flag : int 0 1 1 1 1 1 1 1 1 1 ...  
## $ COHORT\_ID : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MATCHED\_TRIAD : int 48 73 20 38 75 37 41 70 16 50 ...  
## $ SEX\_NUM : int 0 0 1 1 1 0 0 0 0 0 ...  
## $ AGE\_YEARS : int 7 7 3 3 11 3 3 5 4 2 ...  
## $ INDEX\_DATE\_SHIFT : Factor w/ 188 levels "1-Feb-15","1-Jul-15",..: 80 140 9 76 33 130 73 54 100 131 ...  
## $ ED\_COUNT : int 0 0 0 0 0 0 0 1 1 0 ...  
## $ ED\_MEDIAN\_MINS : num 0 0 0 0 0 0 0 135 193 0 ...  
## $ ED\_MEDIAN\_HOURS : num 0 0 0 0 0 0 0 2.25 3.22 0 ...  
## $ ED\_MEDIAN\_DAYS : num 0 0 0 0 0 0 0 0.09 0.13 0 ...  
## $ UC\_COUNT : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ UC\_MEDIAN\_MINS : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ UC\_MEDIAN\_HOURS : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ UC\_MEDIAN\_DAYS : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ICU\_COUNT : int 0 0 1 1 0 0 0 0 0 0 ...  
## $ ICU\_MEDIAN\_MINS : num 0 0 13753 21925 0 ...  
## $ ICU\_MEDIAN\_HOURS : num 0 0 229 365 0 ...  
## $ ICU\_MEDIAN\_DAYS : num 0 0 9.55 15.23 0 ...  
## $ IP\_COUNT : int 0 1 0 0 1 1 0 4 2 3 ...  
## $ IP\_MEDIAN\_MINS : num 0 882 0 0 2156 ...  
## $ IP\_MEDIAN\_HOURS : num 0 14.7 0 0 35.9 ...  
## $ IP\_MEDIAN\_DAYS : num 0 0.61 0 0 1.5 1.17 0 1.97 0.49 1.08 ...  
## $ ANNUAL\_AIR : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ANNUAL\_GROUND : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ANNUAL\_WALK : int 0 1 0 0 1 1 0 4 3 3 ...  
## $ PROCEDURE\_COUNT : int 0 5 7 14 1 2 0 9 0 3 ...  
## $ OTHER\_OUTPATIENT\_COUNT : int 1 11 12 14 6 6 4 22 10 9 ...  
## $ BEHAVIORAL\_COUNT : int 0 0 0 0 0 2 0 31 0 0 ...  
## $ TOTAL\_OUTPATIENT\_COUNT : int 1 11 12 14 6 8 4 53 10 9 ...  
## $ LAB\_COUNT : int 0 4 23 25 9 0 8 14 6 20 ...  
## $ MED\_GIVEN\_OR\_PRESCRIBED: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ PRESCRIPTION\_COUNT : int 1 20 144 256 23 13 10 121 25 81 ...  
## $ DISTANCE : int 3 3 2 3 3 2 2 2 0 3 ...  
## $ INSURANCE : int 1 1 1 0 0 0 1 1 0 0 ...  
## $ TOTAL\_CHARGES : num 140 14317 426846 2361197 24002 ...  
## $ TOTAL\_COST : num 46.3 4724.7 140859.3 779195 7920.6 ...  
## $ TOTAL\_OUTPATIENT\_CHARGE: num 140 14213 321422 2164491 3797 ...  
## $ TOTAL\_OUTPATIENT\_COST : num 46.3 4690.3 106069.4 714282 1253.1 ...  
## $ TOTAL\_INPATIENT\_CHARGE : num 0 104 105424 196706 20205 ...  
## $ TOTAL\_INPATIENT\_COST : num 0 34.3 34789.9 64913 6667.6 ...  
## $ X : logi NA NA NA NA NA NA ...

* The table is in a 'one-patient-one-record' form. The top 6 observations above indicate the structure of the data that will be used for the prediction modelling.

***1c : Discard unnecessary variables in prediction and make dummy variables for those that are necessary and non numeric***

***1d : Missing data check***

|  |  |
| --- | --- |
| Data | Values |
| Total observations | 189 |
| Observation without missing | 188 |
| variables | 58 |

* 1 person with missing insurance information excluded

***1e : Combine the severe epilepsy group***

***1f : Remove variables that produce Near Zero or Zero Variance***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | freqRatio | percentUnique | zeroVar | nzv |
| UC\_COUNT | 25.71429 | 1.5873016 | FALSE | TRUE |
| UC\_MEDIAN\_MINS | 90.00000 | 4.7619048 | FALSE | TRUE |
| UC\_MEDIAN\_HOURS | 90.00000 | 4.7619048 | FALSE | TRUE |
| UC\_MEDIAN\_DAYS | 90.00000 | 3.7037037 | FALSE | TRUE |
| ANNUAL\_AIR | 46.25000 | 1.0582011 | FALSE | TRUE |
| ANNUAL\_GROUND | 187.00000 | 1.5873016 | FALSE | TRUE |
| MED\_GIVEN\_OR\_PRESCRIBED | 0.00000 | 0.5291005 | TRUE | TRUE |
| INSURANCE.2 | 188.00000 | 1.0582011 | FALSE | TRUE |
| INSURANCE.3 | 36.80000 | 1.0582011 | FALSE | TRUE |
| Data | Values |
| observations | 189 |
| variables | 51 |

* After excluding the above variables, there are now 51 variables in the dataset.

***1g***: Split data into training and test.

* Here we split the data into 60/40 (Training/Testing).

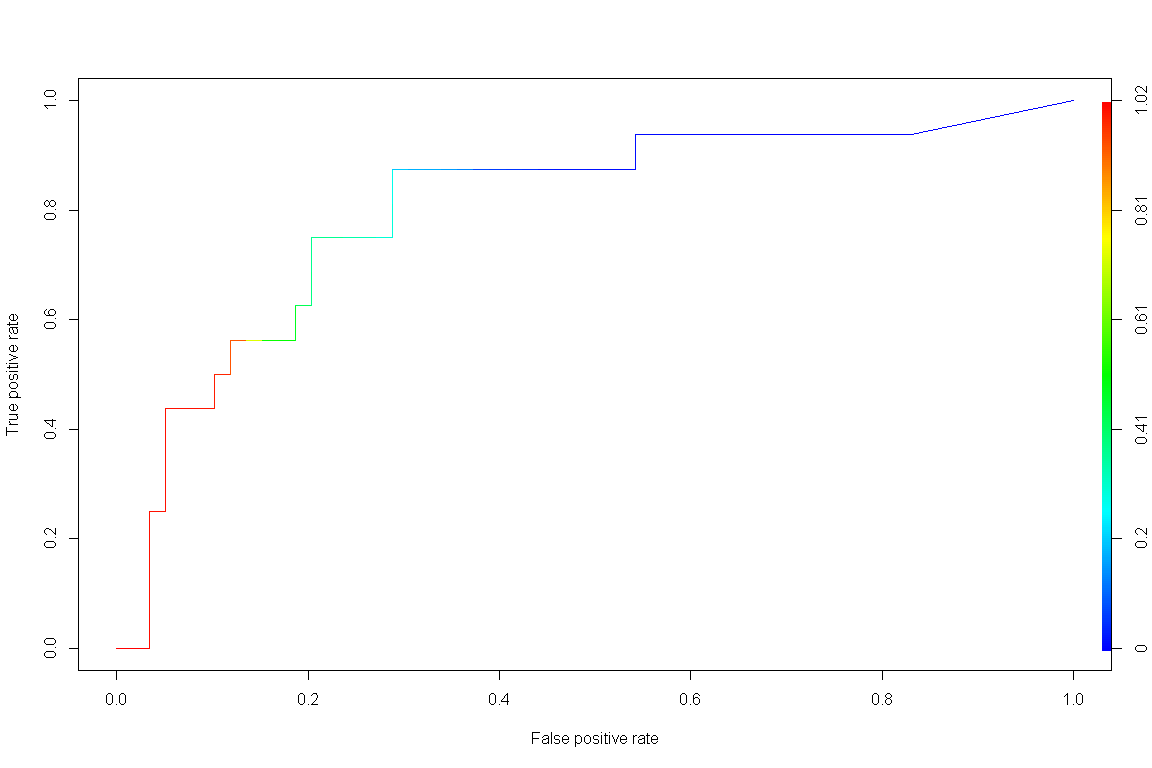
##### Step 2: Logistic Regression Model

***2A: GLM model without variable selection***

Table 1. 2x2 Contingency table without model selection

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | True Status |
|  |  | Severe | Non Severe |
| Predicted Status | Severe | 49 | 7 |
|  | Non Severe | 10 | 9 |

* The initial model used for logistic regression includes the following variables as the predictors: frequency of midazolamhcl prescribed, frequency of DiazePAM prescribed, frequency of Sodium Chloride prescribed, frequency of divalproex prescribed, frequency of Levetiracetam prescribed, count of # Clobazam prescribed, Count of total prescription claims, diagnosis of cardio-vascular chronic comorbid conditions, diagnosis of respiratory chronic comorbid conditions, diagnosis of tech\_dep chronic comorbid conditions, diagnosis of neuro-muscuclar chronic comorbid conditions, diagnosis of congenial genetic chronic comorbid conditions, Total # of chronic comorbid conditions, Count of total out-patient claims, Count of total inpatient claims, Count of total ICU claims, Count of total emergency department visit claims, Count of # of procedure claims, Count of # of laboratory claims, gender, and age.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***sensitivity*** of the logistic regression model is 0.8305085.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***specificity*** of the logistic regression model is 0.5625.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***positive predictive value*** of the logistic regression model is 0.875.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***negative predictive value*** of the logistic regression model is 0.4736842.
* The ***accuracy*** of the model in terms of correctly classifying the outcome is 0.7733333 and the 2x2 contingency table indicates that there are only out of the misclassified.



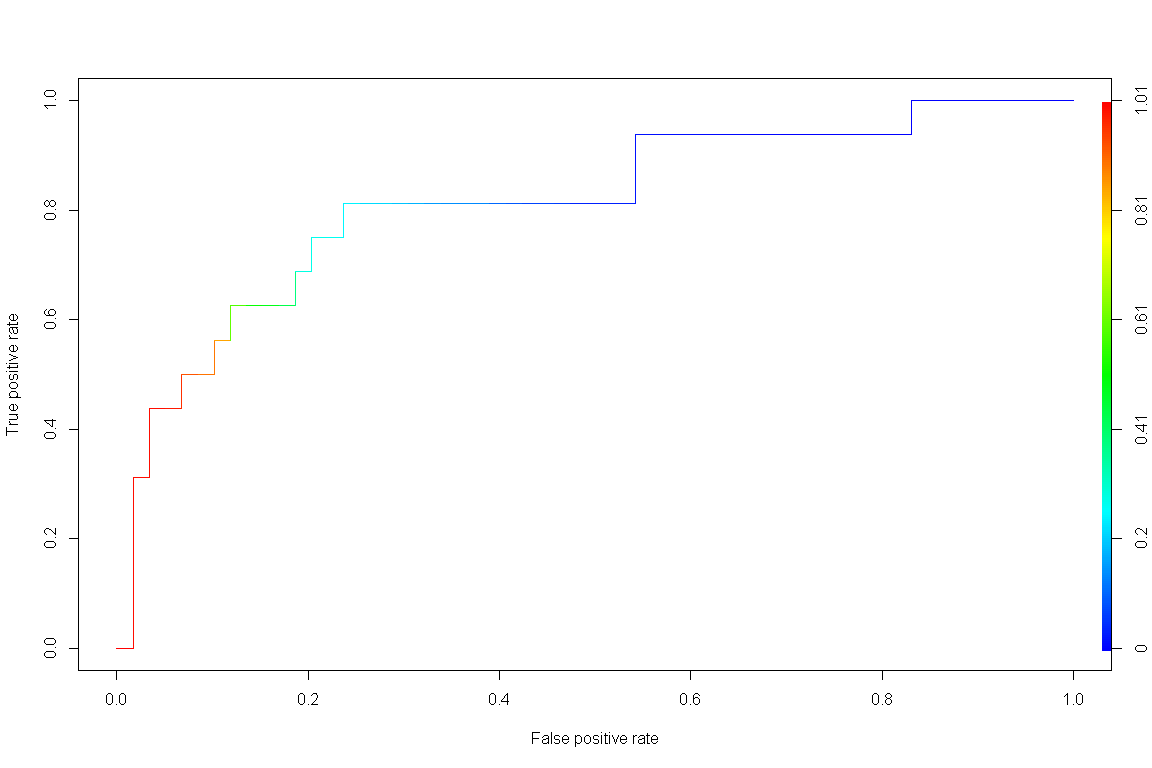
## [1] "AUC :" "0.804025423728814"

* The Area Under the Curve (AUC) of the ROC curve is 0.8040254, which indicates that the probability of the model correctly identifying Severe vs. Non severe is 0.8040254.

***2B: Evaluation of GLM model with Stepwise Variable Selection***:

Table 2. 2x2 contingency table with new model

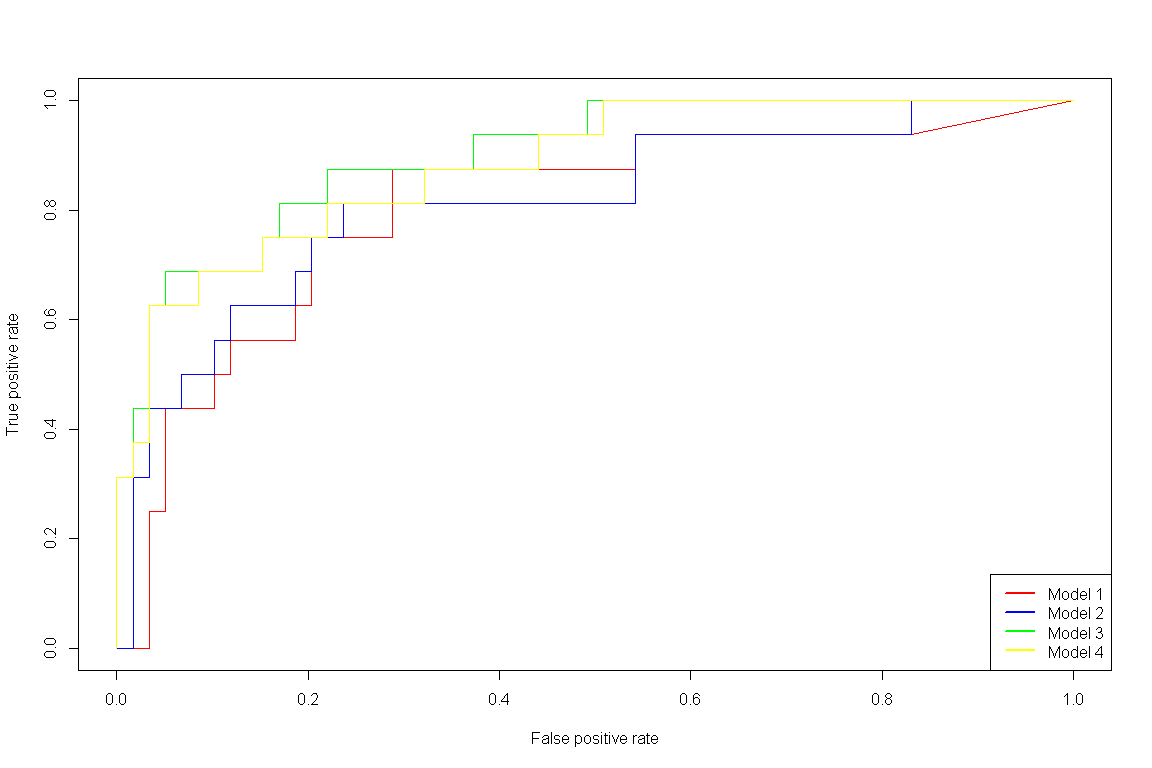
|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | True Status |
|  |  | Severe | Non Severe |
| Predicted Status | Severe | 51 | 6 |
|  | Non Severe | 8 | 10 |



## [1] "AUC :" "0.813559322033898"

* Using the new model created with stepwiseAIC variable selection, given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***sensitivity*** of the logistic regression model is 0.8644068.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***specificity*** of the logistic regression model is 0.625.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***positive predictive value*** of the logistic regression model is 0.8947368.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***negative predictive value*** of the logistic regression model is 0.5555556.
* The ***accuracy*** of the model in terms of correctly classifying the outcome is 0.8133333 and the 2x2 contingency table indicates that there are only out of the misclassified.
* The new model includes variables COHORT3, freq.LevETIRAcetam, cvd\_ccc.0, neuromusc\_ccc.0, TOTAL\_OUTPATIENT\_COUNT, congeni\_genetic\_ccc.0, num\_ccc, freq.CloBAZam, ICU\_COUNT, PROCEDURE\_COUNT which are frequency of Levetiracetam prescribed, Count of # of cardio-vascular chronic comorbid conditions, Count of # of neuro-muscuclar chronic comorbid conditions, Count of total oupatient claims, Count of # of congenial genetic chronic comorbid conditions, Total # of chronic comorbid conditions, frequency of Clobazam prescribed, Count of # of ICU claims, and Count of # of procedure claims.

***2C: Visual Comparison of Model***



* Based on the AUC statistics, the new model (stepwiseAIC using train/test split) yields a slightly higher ***AUC*** of 0.8135593 vs. 0.8040254. The 2 additional new models created by using all of the data (no test/train split with/without the total charges variable included) has an AUC of 0.9004237 and 0.8824153. One thing to note is that Model 1 and Model 2 were both trained using the training data that was created with a 60/40 split whereas Model 3 and Model 4 was built using all of the data thus the AUC will be higher for the latter models.

***2D: Regression parameters***

|  |
| --- |
|  |
|  | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| (Intercept) | 0.37791 (3.77468) | 1.03149 (2.12267) | -0.86949 (0.80795) | -1.11149 (0.79080) |
| freq.midazolamhcl | -0.04655 (0.40664) |  |  |  |
| freq.LevETIRAcetam | 0.40979 (0.26952) | 0.36928 (0.20475) | 0.24049 (0.12204)\* | 0.20604 (0.11680) |
| freq.divalproex | -0.18327 (0.28351) |  |  |  |
| freq.nacl | 0.71794 (1.38071) |  |  |  |
| PRESCRIPTION\_COUNT | -0.05892 (0.08161) |  | -0.10590 (0.03454)\*\* | -0.07711 (0.02773)\*\* |
| freq.DiazePAM | 0.05963 (0.57424) |  | -0.56058 (0.29061) | -0.51518 (0.31688) |
| SEX\_NUM.0 | -0.03547 (0.92900) |  |  |  |
| respiratory\_ccc.0 | 0.23825 (1.40980) |  |  |  |
| cvd\_ccc.0 | -2.97764 (2.09129) | -2.68991 (1.71631) | -1.83994 (0.73518)\* | -1.81316 (0.73696)\* |
| neuromusc\_ccc.0 | 25.54457 (13293.59588) | 20.95812 (6117.56287) | 3.95280 (1.10560)\*\*\* | 4.07301 (1.11314)\*\*\* |
| TOTAL\_OUTPATIENT\_COUNT | -0.35184 (0.18067) | -0.43031 (0.15006)\*\* | -0.14120 (0.06438)\* | -0.15940 (0.06573)\* |
| congeni\_genetic\_ccc.0 | 2.31824 (1.73698) | 1.83639 (1.32130) | 2.01498 (0.77073)\*\* | 2.07514 (0.77678)\*\* |
| tech\_dep\_ccc.0 | -0.03046 (1.09169) |  |  |  |
| num\_ccc | -0.66620 (0.68988) | -0.67208 (0.44744) |  |  |
| AGE\_YEARS | 0.03686 (0.08974) |  |  |  |
| freq.CloBAZam | -0.19789 (0.22347) | -0.22154 (0.14571) |  |  |
| ED\_COUNT | -1.14393 (1.31446) |  |  |  |
| ICU\_COUNT | -54.00276 (2119.56534) | -28.88157 (2084.52608) |  |  |
| IP\_COUNT | -0.38926 (1.25838) |  |  |  |
| PROCEDURE\_COUNT | 1.67609 (0.59054)\*\* | 1.46361 (0.42820)\*\*\* | 0.83833 (0.21618)\*\*\* | 0.85261 (0.21165)\*\*\* |
| LAB\_COUNT | 0.00107 (0.17970) |  |  |  |
| TOTAL\_CHARGES | 0.00001 (0.00005) |  | 3.232e-06  (1.576e-06)\* |  |
| AIC | 97.77998 | 75.40105 | 128.03439 | 128.58510 |
| BIC | 160.71254 | 102.76304 | 160.45186 | 157.76083 |
| Log Likelihood | -25.88999 | -27.70053 | -54.01720 | -55.29255 |
| Deviance | 51.77998 | 55.40105 | 108.03439 | 110.58510 |
| Num. obs. | 114 | 114 | 189 | 189 |
| ***\*\*\*p < 0.001, \*\*****p < 0.01, \**p < 0.05 | | | | |
|  |

* Interpretations below are using Model 3 (Regression Model built using all observations)
* Based on the regression parameters estimated, frequency of Levetiracetam, prescription count, diagnosis of cardiovascular disease chronic comorbid condition, diagnosis of neuromuscular chronic comorbid condition, count of outpatient claims, diagnosis of congenital genetic chronic comorbid condition, procedure count, and total charges had a statistically significant association with the severity of Dravet (p<0.05 for all).
* There is a higher probability of being classified as Non severe epilepsy for increased prescription count, increased total outpatient count, and diagnosis of cardio vascular chronic comorbid condition.
* There is a higher probability of being classified as severe epilepsy for increased frequency of Levetiracetam prescription fill, diagnosis of neuromuscular chronic comorbid condition, diagnosis of congenital genetic chronic comorbid condition, and increased # of procedure count.
* The odds of being classified as Non severe epileptic (i.e. Childhood Absence Epilepsy) decreases by a factor of 0.510 for a one unit increase in the # of outpatient claims.
* The odds of being classified as Non severe epileptic (i.e. Childhood Absence Epilepsy) decreases by a factor of 0.650 for having a diagnosis of cardiovascular chronic comorbid condition.
* The odds of being classified as Non severe epileptic (i.e. Childhood Absence Epilepsy) increases by a factor of 0.801 for having a diagnosis of congenital genetic chronic comorbid condition.

***Probability of classified as Non Severe Epilepsy: adjusting Prescription claims***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Levetiracetam** | | **CVD** | **Neuro mus** | **Outpat count** | | **Congen. CCC** | | **Tot CCC** | | **CloBAZam** | | **ICU** | **Proc count** | | | **DiazePAM** | | **Charges** | | **Presc.count** | | **P(NS)** |
| 1.93 | 0.703 | | 0.0846 | 12.10 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 2 | | 0.582 | |
| 1.93 | 0.703 | | 0.0846 | 12.10 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 5 | | 0.503 | |
| 1.93 | 0.703 | | 0.0846 | 12.10 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 10 | | 0.374 | |
| 1.93 | 0.703 | | 0.0846 | 12.10 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 15 | | 0.260 | |

* After setting all variables to their mean values except for # of annual prescription claims, adjusting the values of the # of prescription claims indicate that the probability of being classified as non-severe epilepsy is 0.582, 0.504, 0.374, and 0.26 as we adjust the # of annual prescription claims to 2, 5, 10, and 15, respectively.

***Probability of classified as Non Severe Epilepsy: Adjusting TOTAL\_OUTPATIENT\_COUNT***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Levetiracetam** | | **CVD** | **Neuro mus** | **Outpat count** | | **Congen. CCC** | | **Tot CCC** | | **CloBAZam** | | **ICU** | **Proc count** | | | **DiazePAM** | | **Charges** | | **Presc.count** | | **P(NS)** |
| 1.93 | 0.703 | | 0.0846 | 1 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 40 | | 0.104 | |
| 1.93 | 0.703 | | 0.0846 | 2 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 40 | | 0.092 | |
| 1.93 | 0.703 | | 0.0846 | 3 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 40 | | 0.080 | |
| 1.93 | 0.703 | | 0.0846 | 4 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 40 | | 0.071 | |
| 1.93 | 0.703 | | 0.0846 | 5 | 0.64 | | 2.86 | | 3.02 | | 0.185 | | | 3.16 | 1.513 | | 161054 | | 40 | | 0.062 | |

* After setting all variables to their mean values except for # of total outpatient claims, adjusting the values of the # of total outpatient claims indicate that the probability of being classified as non-severe epilepsy is 0.105, 0.092, 0.081, 0.071, and 0.062 as we adjust the # of outpatient claims to 1,2,3,4, and 5, respectively.

| Table 1. Baseline Demographic/Clinical Characteristics of Study Population | | |
| --- | --- | --- |
|  | **level** | **Overall** |
| **n** |  | 191 |
| **AGE\_YEARS (mean (sd))** |  | 8.92 (5.58) |
| **SEX\_NUM (%)** | 0 | 95 (49.7) |
|  | 1 | 96 (50.3) |
| **freq.LevETIRAcetam (mean (sd))** |  | 1.92 (4.39) |
| **PRESCRIPTION\_COUNT (mean (sd))** |  | 39.94 (84.86) |
| **freq.DiazePAM (mean (sd))** |  | 1.50 (3.88) |
| **cvd\_ccc (%)** | 0 | 135 (70.7) |
|  | 1 | 56 (29.3) |
| **neuromusc\_ccc (%)** | 0 | 17 ( 8.9) |
|  | 1 | 174 (91.1) |
| **TOTAL\_OUTPATIENT\_COUNT (mean (sd))** |  | 12.08 (14.23) |
| **congeni\_genetic\_ccc (%)** | 0 | 122 (63.9) |
|  | 1 | 69 (36.1) |
| **PROCEDURE\_COUNT (mean (sd))** |  | 3.18 (5.12) |
| **TOTAL\_CHARGES (mean (sd))** |  | 159872.21 (1003105.40) |
| **freq.CloBAZam (mean (sd))** |  | 2.99 (4.86) |
| **ICU\_COUNT (mean (sd))** |  | 0.18 (0.54) |

***2F***: Histogram to see distribution of variables within the Dravet Cohort used in Colorado

