Modeling the Probability of an Insured Person Being Hospitalized >3 Times

Introduction:

The motivation behind this project is to find the model with the best prediction accuracy that correctly classifies Blue Cross Blue Shield of RI insured members as being hospitalized >3 times. This type of modelling is useful because it can be applied on data of new members to classify and rank them as being low, medium, or high risk in order to appropriately establish an individualized pricing scheme that reflects the amount of risk they carry. Even more specifically, their individual probabilities of being hospitalized >3 times can be extracted and a direct ranking can be developed for internal purposes.

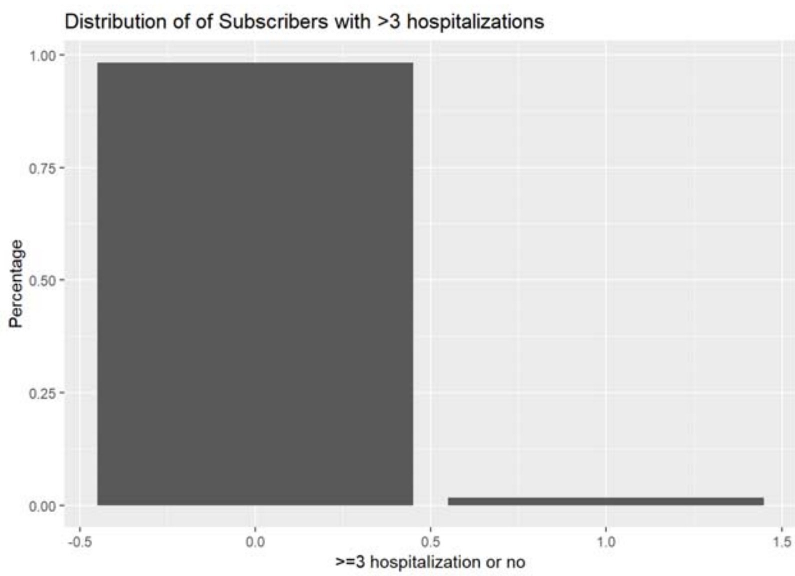
To begin this analysis, the data set “ed_2017.sas7bdat” from Blue Cross Blue Shield RI was used. This data set reflects 2017 data and has 155,769 rows of data and 61 columns. Each row reflects the data of an individual and there are 60 predictand columns and 1 predictor column that designates a “1” if the individual was hospitalized 3 or more times and “0” if otherwise. The table in the appendix includes a description of *all* the variables for clarity. However, in our analysis, we used a subset of these variables to build the models.

Methods:

The first step of modeling taken was to first establish a high-level understanding of the data. We did this by computing tables of the variables to better understand the distribution of the variables. Included in the appendix are the distributions of all of the variables, but in this section we will highlight the distribution of the predictor because of its significant impact on our procedure for modeling.

Distribution of outcome variable:

The outcome variable is called “OP_ER_Cnt_Target.” This variable is “0” if a person is hospitalized less than 3 times and “1” otherwise. The distribution is as follows:



This outcome variable is extremely unbalanced. More specifically, the percentage of insured individuals who have been hospitalized 3 or more times is comparatively low. This means typical regression with a training set comprised of 80% of the original data and a testing set comprised of 20% of the original data will yield poor results since there will not be enough people with 3 or more hospitalizations to train the model properly. There are several paths within the realm of non-penalized regression to solve this problem:

1. A training set that over samples the minority class
2. A training set that under samples the majority class
3. A training set that does both
4. A training set built with synthetic data based on the present data.

We will build models using each of these training data sets to see which ones give us the best results.

Oversampling Training Data set

For the oversampling method, the training set is built by sampling *with replacement* of the minority class alone. This means that many rows will be repeated until we have a data set that we feel is large enough to train on. An advantage of this method is that no data is thrown away. All of the rows in the original training set will be there, but some will be repeated. A disadvantage of this method is that it often leads to overfitting since many rows are typically repeated. This means that the predictive power of the model is typically not as high as the same model with a differently designed training set.

Under sampling Training Data set

For the under sampling method, the training set is built by sampling a set number of majority class rows such that the resulting training set is at most twice the size of the minority class. This leads to a loss of information in the model since the number of 1's is the limiting factor in terms of the size of the data set. An advantage of this training set is that it generally leads to better predictions in the model because it does not overfit. However, a significant amount of data can be lost, which also affects general model performance since there is less information included to inform the model.

Under sampling the Majority Class and Oversampling the Minority Class

This method is in theory the best of both worlds. The majority class is under sampled and the minority class is over sampled. The minority class will be over sampled until the specified size of the desired data set.

Regenerated Training Data

This methods creates synthetic data that is based on the real data. It build artificial examples from the classes based on the smoothed bootstrapping approach.

Determining Tuning Parameters:

To build the aforementioned training data sets, there were 2 tuning parameters that needed to be adjusted. The first tuning parameter is the percentage of the data set that we wanted to be 1's. We could have set this to be 50/50, but we chose to adjust it to whatever gives the best prediction results in the modeling. The second tuning parameter that we adjusted was the boundaries for setting the predicted probability to 0 or 1. If the predicted probability of an individual is 0.45 should they assigned a 0 or 1? This is a tricky question because often the 0.5 cutoff does not actually give the best predictions on test data. Therefore, the predicted probability boundary was also tuned to give the best prediction power to the model.

To do this algorithmically, a function was built that outputs a 3-dimensional matrix. Another function uses this matrix then automatically selects the top 5 highest average accuracy across both classes and then selects the the parameters that give the minimum absolute difference in accuracy between each class. This selection of the threshold and the percent of each class in the data set is critical. We assumed here that we would prefer similar accuracies between the classes; however, this is debatable. A further discussion of this will be in the conclusion.

After the parameters were selected, this was run 10 times. We also looked at the best parameters to see if the mode would be a more appropriate choice compared to the mean. In all cases, we identified and used the mode from this output to develop the training sets.

Data Cleaning:

There were 749 NA's in the data set. These entire rows were removed. In addition, there were only 4 entries that had MEMBER_RELATIONSHIP_DESC == "OTHER RELATIONSHIP." These were also removed because it led to errors in the modeling when the testing set by chance had all 4 entries and the training set did not have one since the model cannot be evaluated on new levels in the testing set.

Models:

There are several different models that we used in this analysis:

1. Logistic Regression
2. Decision Tree
3. Random Forest

These are all very different models so the motivation behind each model will be detailed in the following paragraphs.

Logistic Regression:

Considering that our outcome variable is binary, this was a very natural place to start since it is a basic classification regression. All of the covariates in our models are shown below for clarity. This basic formula did not change over different training sets or modeling techniques.

```
OP_ER_Cnt_Target ~ Age + Gender + MEMBER_RELATIONSHIP_DESC + Product + Business_Segment + Hypertension + Hyperlipid + LowBac
kPain + Diabetes + IschemicHD + Asthma + COPD + CHF + Cancer + HIV_AIDS + Depression + SubstanceAbuse + Schizophenia + Perso
nalityDisorder + Bipolar + Dementia + RUB + IP_Total_Cnt + Total_Allowed + OP_ER_Cnt_previous + ESRD_Flag + Fall_Flag + PCMH
_Info
```

The following subsections will provide the results for each of the models.

Comparison of the Models with Corrected Training Sets and the Standard 80/20 Train/Test

	80/20 Split Model			**Both Over and Under**			**Synthetic Data**			**Oversampling Model**			**Undersampling model**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
Intercept	0.00	0.00 – 0.00	**<0.001**	0.06	0.06 – 0.07	**<0.001**	0.26	0.24 – 0.29	**<0.001**	0.10	0.09 – 0.11	**<0.001**	0.10	0.06 – 0.17	**<0.001**
Age	1.00	0.99 – 1.00	0.092	0.99	0.99 – 1.00	**<0.001**	1.00	1.00 – 1.00	**<0.001**	0.99	0.99 – 0.99	**<0.001**	0.99	0.99 – 1.00	**0.016**
Gender: M	0.83	0.75 – 0.91	**<0.001**	0.84	0.82 – 0.87	**<0.001**	0.83	0.81 – 0.86	**<0.001**	0.85	0.83 – 0.87	**<0.001**	0.86	0.75 – 0.99	**0.040**
Member_Relationship: Employee	0.57	0.46 – 0.70	**<0.001**	0.61	0.57 – 0.65	**<0.001**	0.50	0.48 – 0.53	**<0.001**	0.61	0.58 – 0.64	**<0.001**	0.58	0.44 – 0.78	**<0.001**

Member_Relationship: Handicapped Dependent	1.48	0.64 – 3.42	0.354	1.64	1.15 – 2.35	**0.007**	1.83	1.35 – 2.47	**<0.001**	1.70	1.32 – 2.18	**<0.001**	4.22	0.69 – 25.80	0.120
Member_Relationship: Life Partner	0.26	0.03 – 2.05	0.203	0.39	0.23 – 0.66	**<0.001**	0.25	0.15 – 0.40	**<0.001**	0.40	0.28 – 0.58	**<0.001**	0.46	0.05 – 4.50	0.502
Member_Relationship: Significant Other	0.78	0.24 – 2.55	0.680	1.14	0.77 – 1.67	0.523	0.94	0.66 – 1.33	0.723	0.95	0.70 – 1.27	0.720	1.75	0.30 – 10.23	0.535
Member_Relationship: Spouse	0.54	0.42 – 0.68	**<0.001**	0.62	0.58 – 0.67	**<0.001**	0.50	0.48 – 0.53	**<0.001**	0.61	0.58 – 0.64	**<0.001**	0.63	0.45 – 0.86	**0.004**
Product: FEP	0.82	0.60 – 1.14	0.247	0.77	0.72 – 0.82	**<0.001**	0.73	0.67 – 0.79	**<0.001**	0.76	0.70 – 0.81	**<0.001**	0.80	0.52 – 1.24	0.321
Product: MedAdvantage	1.93	1.47 – 2.54	**<0.001**	2.30	2.18 – 2.42	**<0.001**	2.21	2.07 – 2.36	**<0.001**	2.25	2.11 – 2.39	**<0.001**	2.18	1.49 – 3.19	**<0.001**
Business_Segment: Large Group	0.99	0.75 – 1.29	0.920				0.97	0.91 – 1.04	0.459	0.96	0.91 – 1.02	0.181	1.01	0.70 – 1.44	0.971
Business_Segment: Self Insured	0.98	0.76 – 1.27	0.886				1.00	0.94 – 1.07	0.888	0.98	0.92 – 1.03	0.372	0.91	0.65 – 1.28	0.596
Business_Segment: Small Group	1.00	0.75 – 1.33	0.973				0.98	0.91 – 1.05	0.563	0.97	0.91 – 1.03	0.330	0.89	0.61 – 1.32	0.568
Hypertension	1.30	1.16 – 1.45	**<0.001**	1.18	1.14 – 1.22	**<0.001**	1.18	1.15 – 1.20	**<0.001**	1.23	1.20 – 1.26	**<0.001**	1.32	1.12 – 1.56	**0.001**
Hyperlipid	0.95	0.85 – 1.07	0.393	0.95	0.92 – 0.99	**0.013**	1.03	1.01 – 1.05	**0.013**	1.00	0.97 – 1.03	0.997	0.90	0.76 – 1.07	0.235
LowBackPain	1.47	1.31 – 1.66	**<0.001**	1.58	1.52 – 1.65	**<0.001**	1.39	1.36 – 1.43	**<0.001**	1.57	1.53 – 1.62	**<0.001**	1.54	1.28 – 1.86	**<0.001**
Diabetes	1.31	1.16 – 1.48	**<0.001**	1.33	1.28 – 1.39	**<0.001**	1.23	1.20 – 1.27	**<0.001**	1.31	1.27 – 1.36	**<0.001**	1.43	1.18 – 1.74	**<0.001**
IschemicHD	1.47	1.27 – 1.70	**<0.001**	1.58	1.50 – 1.66	**<0.001**	1.37	1.33 – 1.42	**<0.001**	1.57	1.50 – 1.63	**<0.001**	1.61	1.27 – 2.05	**<0.001**
Asthma	1.27	1.11 – 1.46	**0.001**	1.27	1.21 – 1.33	**<0.001**	1.21	1.17 – 1.25	**<0.001**	1.26	1.21 – 1.30	**<0.001**	1.29	1.04 – 1.61	**0.023**
COPD	1.15	0.96 – 1.38	0.119	1.05	0.98 – 1.13	0.140	1.23	1.18 – 1.29	**<0.001**	1.03	0.98 – 1.09	0.254	0.97	0.71 – 1.34	0.874
CHF	1.28	1.05 – 1.57	**0.016**	1.64	1.51 – 1.78	**<0.001**	1.48	1.41 – 1.56	**<0.001**	1.81	1.70 – 1.93	**<0.001**	2.16	1.45 – 3.24	**<0.001**
Cancer	0.80	0.68 – 0.93	**0.005**	0.82	0.78 – 0.86	**<0.001**	0.94	0.91 – 0.98	**0.001**	0.76	0.73 – 0.79	**<0.001**	0.81	0.64 – 1.03	0.090
HIV_AIDS	2.46	1.19 – 5.09	**0.015**	1.96	1.52 – 2.53	**<0.001**	1.58	1.32 – 1.90	**<0.001**	1.77	1.46 – 2.16	**<0.001**	1.95	0.55 – 6.89	0.300
Depression	1.42	1.25 – 1.62	**<0.001**	1.50	1.44 – 1.57	**<0.001**	1.43	1.39 – 1.48	**<0.001**	1.45	1.41 – 1.50	**<0.001**	1.44	1.17 – 1.77	**<0.001**
SubstanceAbuse	1.97	1.63 – 2.38	**<0.001**	2.33	2.16 – 2.51	**<0.001**	1.83	1.75 – 1.92	**<0.001**	2.38	2.25 – 2.52	**<0.001**	2.55	1.80 – 3.61	**<0.001**
Schizophenia	2.37	1.64 – 3.41	**<0.001**	2.59	2.20 – 3.04	**<0.001**	1.66	1.50 – 1.83	**<0.001**	2.76	2.43 – 3.13	**<0.001**	2.98	1.32 – 6.72	**0.008**
PersonalityDisorder	1.76	1.03 – 3.03	**0.040**	1.41	1.09 – 1.82	**0.008**	1.53	1.33 – 1.76	**<0.001**	1.37	1.11 – 1.67	**0.003**	0.75	0.27 – 2.09	0.581
Bipolar	1.39	1.06 – 1.81	**0.017**	1.49	1.34 – 1.66	**<0.001**	1.58	1.48 – 1.69	**<0.001**	1.79	1.65 – 1.94	**<0.001**	1.82	1.10 – 3.01	**0.019**
Dementia	1.47	1.14 – 1.90	**0.003**	2.05	1.85 – 2.28	**<0.001**	1.40	1.31 – 1.50	**<0.001**	1.93	1.78 – 2.10	**<0.001**	2.16	1.30 – 3.58	**0.003**
RUB	1.64	1.52 – 1.77	**<0.001**	1.59	1.55 – 1.63	**<0.001**	1.45	1.42 – 1.47	**<0.001**	1.58	1.55 – 1.62	**<0.001**	1.60	1.41 – 1.80	**<0.001**
IP_Total_Cnt	1.08	1.02 – 1.14	**0.008**	1.06	1.03 – 1.09	**<0.001**	1.08	1.07 – 1.10	**<0.001**	1.05	1.03 – 1.07	**<0.001**	0.94	0.84 – 1.06	0.323

Total_Allowed	1.00	1.00 – 1.00	**0.001**	1.00	1.00 – 1.00	**<0.001**	1.00	1.00 – 1.00	**<0.001**	1.00	1.00 – 1.00	**<0.001**	1.00	1.00 – 1.00	**<0.001**
OP_ER_Cnt_previous	1.62	1.57 – 1.67	**<0.001**	2.13	2.10 – 2.17	**<0.001**	1.22	1.21 – 1.22	**<0.001**	2.19	2.16 – 2.22	**<0.001**	2.22	2.04 – 2.41	**<0.001**
ESRD_Flag: Y	1.19	0.64 – 2.21	0.587	1.17	0.87 – 1.58	0.309	1.33	1.02 – 1.74	**0.033**	0.94	0.75 – 1.18	0.613			
Fall_Flag: Y	1.65	1.40 – 1.94	**<0.001**	1.92	1.80 – 2.05	**<0.001**	2.47	2.32 – 2.62	**<0.001**	1.98	1.88 – 2.08	**<0.001**			
PCMH_Info: No PCP	1.16	1.00 – 1.34	0.054	1.16	1.10 – 1.21	**<0.001**	1.14	1.10 – 1.19	**<0.001**	1.17	1.13 – 1.21	**<0.001**			
PCMH_Info: Has PCP	0.87	0.79 – 0.97	**0.009**	0.83	0.81 – 0.86	**<0.001**	0.85	0.83 – 0.87	**<0.001**	0.86	0.84 – 0.88	**<0.001**			
Observations	124016			124016			124016			203177			5244		
Tjur's R^2	0.082			0.305			0.218			0.313			0.308		

Above in bold we can see all of the exponentiated coefficients and the exponentiated confidence intervals with the significant ones highlighted in bold. We can see for example in the oversampling method that those who abuse substances have 2.38 times the odds of being hospitalized 3 or more times compared to someone who does not abuse substances.

At first glance, the results from the logistic regression applied to different training sets seem very similar because the bulk of the statistically significant coefficients are the same across models. However, these models differ in their R^2 values. We can see that the under sampling training set and the oversampling training set captures much more of the variance compared to the original training 80/20 training set.

Picking out some of the more interesting coefficients, we can see ESRD (End Stage Renal Disease) and COPD are not significant across the model for all training sets except for with the oversampling training set. It is logical that most of these people would not spend 3 or more days in a hospital because of the advances in modern medicine that have significantly improved the prognosis. If we were instead doing this research in the 1980's, these covariates would likely be very significant in determining the length of stay.

Another interesting covariate would be gender. Being male actually decreases the odds of staying in a hospital 3 or more days. This was true across all training sets and the coefficients were all less than 1 within a 95% confidence interval.

Lastly, we can see from the table above that the coefficients for age were all essentially 1, and they were significant across all training sets except for the 80/20 training set. This implies that age does *not* have an impact on the odds of staying 3 or more days in a hospital. This find was quite surprising because we expected older people to have longer stays than younger people.

Model Performance and Goodness of Fit Comparison:

	80/20	Both Over and Under	Synthetic Data	Oversampling Model	Undersampling model
McFadden Value:	0.1630199	0.250521	0.1723772	0.2526101	0.2485301
Hosmer-Lemeshow p-value:	6.070699510^{-13}	0	0	0	8.059934910^{-10}

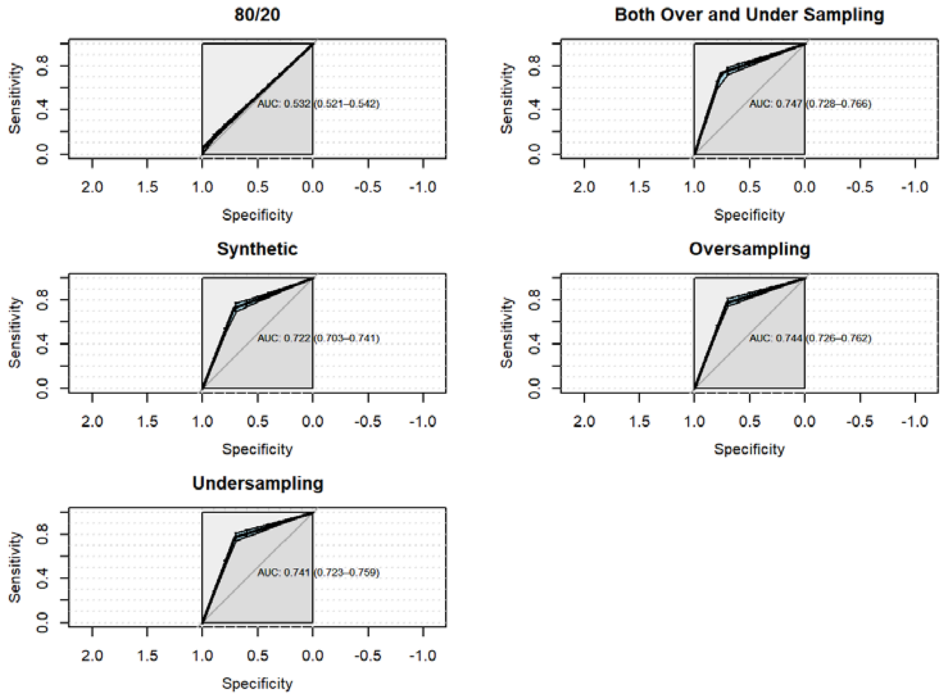
Looking at the model performance we can see that we have a poor fit across all training sets according to both the McFadden pseudo R^2 and the hoslem test. This suggests that our model does not accurately characterize the data and could be improved. However, we would like to note that the R^2 does improve using our constructed training sets compared to just the 80/20 training set.

Prediction Power:

	Average Accuracy	Accuracy for Class 1	Misclassification for Class 1	Accuracy for Class 0	Misclassification for Class 0
80/20	98.3002193	6.4150943	93.5849057	99.8982739	0.1017261
Both Over and Under	76.3030577	73.0188679	26.9811321	76.3601759	23.6398241
Synthetic Data	71.8068636	72.6415094	27.3584906	71.7923476	28.2076524
Oversampling Model	71.3230551	77.5471698	22.4528302	71.2148061	28.7851939
Undersampling Model	71.2037156	77.1698113	22.8301887	71.0999541	28.9000459

Most importantly, we were interested in the prediction power of the model. We have excellent accuracy for Target = 0, but very poor accuracy for Target = 1 for the original 80/20 training set. This was as expected because we have a very unbalanced data set, which is why we developed special training sets. For the other training sets we get between 71-78% accuracy across all models and between classes. This is a dramatic improvement from the original model since the misclassification rate of class "1" was 94% in it.

We can also see in the ROC curves for the corresponding training sets a dramatic improvement from the 80/20 original training set. The area under the curve (AUC) for the 80/20 training set was 0.532, which implies the model is as good as a coin flip. The AUC increased dramatically for the others to reach 0.747 for the hybrid over/under sampling training sets, 0.744 for the oversampling training set, 0.741 for the undersampling training set, and 0.722 for the synthetic data.

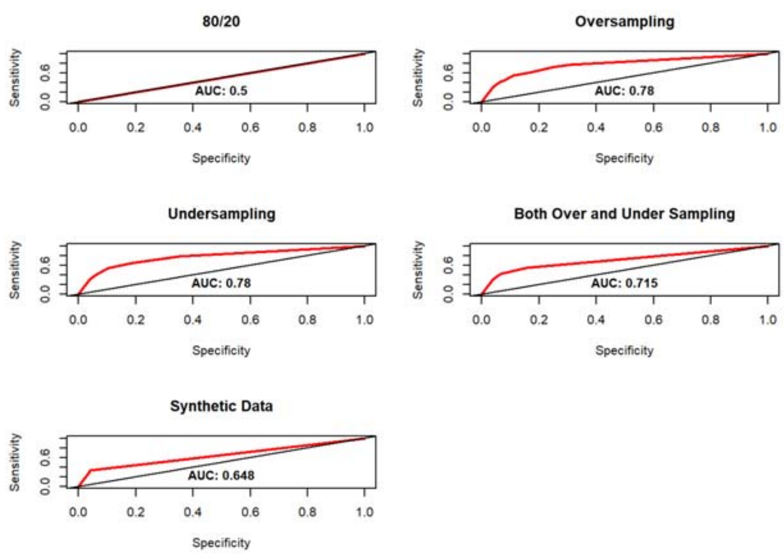


In summary, We have decent accuracy with this model, but there is still room for improvement since we misclassify about 25% for each class. The best training set to use with this model is the oversampling training set which correctly classifies class 1 78% of the time and class 0 71% of the time. It has the highest McFadden value and Tjur's R^2 , and it also has an AUC of 0.744, which is quite high comparatively although very close to the AUC of the hybrid training set. It was surprising for the oversampling method to outperform the undersampling method with respect to the misclassification rate because we expected it to overfit initially. Even more surprising is that the AUC of undersampling and oversampling are very similar, which means their predictive abilities should be similar, which they are.

We continued to try other modeling techniques in the hopes that one of them will have much higher performance.

Decision Tree

We followed the same pattern as with logistic regression and built a decision tree for all of the corresponding training sets. The first table shown below highlights the error rates on the training set for the models, and the plot below shows all of the ROC curves for the different training sets.



Looking at the ROC curves, it can be seen that the decision tree performs much better for oversampling and undersampling training sets, but it performs worse on the others. The greatest AUC is 0.78, which is a noticeable improvement from the logistic regression.

The minimum of the error across all splits for each model:

	CP	nsplit	rel error	xerror	xstd	Cross Validated Error Rate
80/20	0	0	1	0	0	0
Both Over and Under	0.01	3	0.7232413	0.7243928	0.0038945	0.2181251

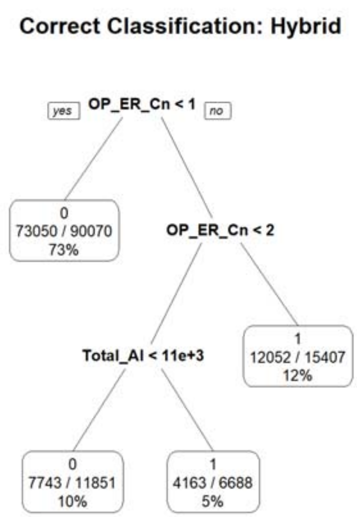
	CP	nsplit	rel error	xerror	xstd	Cross Validated Error Rate
Synthetic Data	0.01	7	0.1856847	0.1864771	0.0016539	0.0929799
Oversampling Model	0.01	7	0.6083692	0.6085662	0.0023806	0.2433642
Undersampling Model	0.01	6	0.6092612	0.6095047	0.0030614	0.2422026

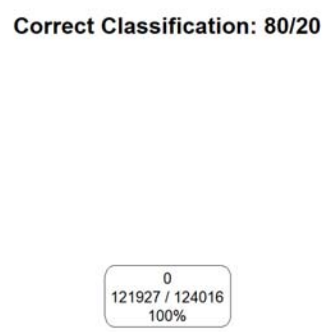
In the table shown above, we notice that the synthetic data has the lowest cross-validated error rate. This was surprising because in logistic regression and in the decision tree it under performed compared to the other training sets. We hypothesize that the decision tree for synthetic data has the potential to overfit and therefore gives good results on the training data but then gives poor prediction results. This will be investigated further in the table below with a pruned tree.

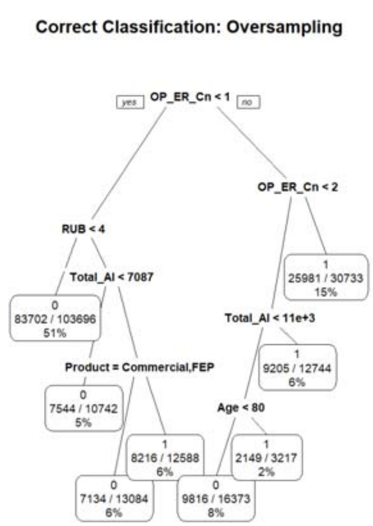
	80/20		Oversampling		Undersampling		Both		Synthetic	
	Pred 0	Pred 1	Pred 0	Pred 1	Pred 0	Pred 1	Pred 0	Pred 1	Pred 0	Pred 1
Actual 0	30474	NA	22903	7571	19767	10707	25737	4737	29199	1275
Actual 1	530	NA	147	383	115	415	238	292	351	179

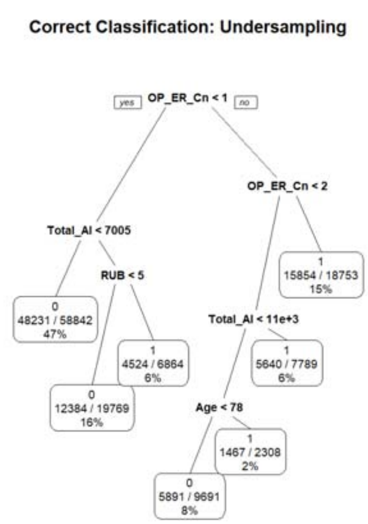
From this table that we obtain after pruning the tree, we first notice that our hypothesis for the synthetic training data set was correct since it does very poorly for the minority class.

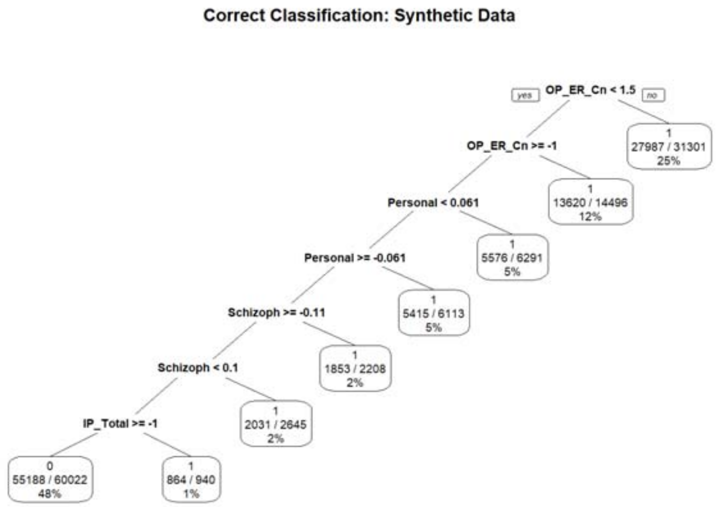
In addition, it is important to be able to visualize the tree. We have a graph of the trees printed below.











From these trees a few themes are salient. The first is that the variable for outpatient ER visits (OP_ER_Cn) and total allowed (Total_Al) are very important in all training sets. The variables RUB and age also play an important role, but they are comparatively smaller.

This model, overall, is not a significant improvement to the logistic regression. The confusion matrix shows that performance for the best decision tree is on par with the logistic regression models. The best model among the decision trees is with the oversampling training set again. The next step of our analysis was to make a more robust version of the decision tree by performing random forest. Since we saw some slight improvement between logistic regression and decision tree models, we were hoping to capitalize on this and enhance the performance utilizing random forest.

Random Forest

The favorite models by far were the random forest generated models because of the ability to tune different parameters. All of these models had misclassification on the order of 1-30% for **both** classes in the training set as shown in the first table below.

	Synthetic			Oversampling			Undersampling			Both			80/20		
	Pred:0	Pred:1	Error	Pred:0	Pred:1	Error	Pred:0	Pred:1	Error	Pred:0	Pred:1	Error	Pred:0	Pred:1	E
0	56459	5377	0.0869558	87290	34637	0.2840798	2237	918	0.2909667	63452	23221	0.2679150	102662	19265	0.1580
1	777	61403	0.0124960	4867	76383	0.0599015	420	1669	0.2010531	4053	33290	0.1085344	695	1394	0.3326

This second table shows the error rates on the test data. It can be seen that the hybrid model has the best prediction accuracy on average between both classes. This model unfortunately was not better than the decision tree or the logistic regression. All of the models are on par with about 75% accuracy in each class with a maximum of 80% depending on the run.

	Average Accuracy	Accuracy for Class 1	Misclassification for Class 1	Accuracy for Class 0	Misclassification for Class 0
80/20	84.2794478	66.7924528	33.2075472	84.5835794	15.4164206
Both Over and Under	73.1550768	78.1132075	21.8867925	73.0688456	26.9311544
Synthetic Data	97.0068378	24.9056604	75.0943396	98.2608125	1.7391875
Oversampling Model	71.9971617	78.3018868	21.6981132	71.8875107	28.1124893
Undersampling Model	70.8069926	79.245283	20.754717	70.660235	29.339765

The last thing that is of interest in the random forest model is the ranking of importances. The table below shows the importance of the variables in all models.

Synthetic		Oversampling		Undersampling		
Covariate	Gini	Covariate	Gini	Covariate	Gini	C
Age	13.7516601	Age	183.3173103	Age	189.4522420	A
Gender	2.1536056	Gender	28.0684713	Gender	28.0962088	G
MEMBER_RELATIONSHIP_DESC	4.1753593	MEMBER_RELATIONSHIP_DESC	31.3517415	MEMBER_RELATIONSHIP_DESC	32.0038817	M
Product	5.7006191	Product	27.9555933	Product	25.5472081	P

Synthetic		Oversampling		Undersampling		
Covariate	**Gini**	**Covariate**	**Gini**	**Covariate**	**Gini**	**C**
Business_Segment	12.8348238	Business_Segment	67.3853027	Business_Segment	66.9202506	E
Hypertension	14.4378717	Hypertension	26.4497233	Hypertension	26.6467692	H
Hyperlipid	13.3788985	Hyperlipid	22.3104807	Hyperlipid	22.3965808	H
LowBackPain	16.2977985	LowBackPain	23.0119341	LowBackPain	22.8937248	L
Diabetes	14.6937169	Diabetes	20.3687909	Diabetes	20.6956653	D
IschemicHD	28.1342096	IschemicHD	16.3804825	IschemicHD	16.9690086	I
Asthma	15.4760728	Asthma	18.8574808	Asthma	19.9609812	A
COPD	36.5191597	COPD	8.8866928	COPD	8.6949683	C
CHF	96.0371867	CHF	8.3712655	CHF	8.7319459	C
Cancer	13.0843258	Cancer	15.3753154	Cancer	15.0650594	C
HIV_AIDS	23.9749968	HIV_AIDS	0.7651241	HIV_AIDS	0.7788557	H
Depression	20.9246890	Depression	21.6697713	Depression	22.0320635	D
SubstanceAbuse	55.8261811	SubstanceAbuse	14.6783925	SubstanceAbuse	14.4449160	S
Schizophenia	142.0167399	Schizophenia	4.1184549	Schizophenia	3.5905531	S
PersonalityDisorder	208.6903256	PersonalityDisorder	0.8614848	PersonalityDisorder	1.3956695	P
Bipolar	51.2056752	Bipolar	7.5992819	Bipolar	7.2217886	B
Dementia	85.8839318	Dementia	5.0759772	Dementia	5.8434903	D
RUB	55.9610754	RUB	83.8483182	RUB	82.1338680	R
IP_Total_Cnt	180.7627503	IP_Total_Cnt	37.6774061	IP_Total_Cnt	37.5682901	I
Total_Allowed	67.0889692	Total_Allowed	259.7307407	Total_Allowed	257.6998490	T
OP_ER_Cnt_previous	314.6972792	OP_ER_Cnt_previous	166.6348410	OP_ER_Cnt_previous	172.2841522	C
ESRD_Flag	0.2174942	ESRD_Flag	0.8738168	ESRD_Flag	1.4328685	E
Fall_Flag	1.9041913	Fall_Flag	15.2392014	Fall_Flag	16.6792679	F
PCMH_Info	4.1703931	PCMH_Info	44.0535104	PCMH_Info	45.4981636	P

Both Over and Under Sampling	
Covariate	**Gini**
Age	200.4086490
Gender	26.5353051
Hypertension	25.2410349
Hyperlipid	23.6832778
LowBackPain	22.6741266
Diabetes	20.3288803
IschemicHD	17.4607880
Asthma	17.9308939
COPD	9.4756936
CHF	9.1112083
Cancer	16.3873906
HIV_AIDS	0.9503957

Both Over and Under Sampling	
Covariate	**Gini**
Depression	21.1654835
SubstanceAbuse	14.4260046
Schizophenia	4.9557934
PersonalityDisorder	0.9549838
Bipolar	7.2768328
Dementia	6.3607557
RUB	85.8865503
IP_Total_Cnt	39.4093713
Total_Allowed	270.5773585
OP_ER_Cnt_previous	178.6709766
ESRD_Flag	0.8485776
Fall_Flag	18.2090400
PCMH_Info	44.0299438

It can be seen that the importances of covariates in the constructed training sets are all similar except for the synthetic data set. This wayward data set placed importance on schizophrenia and personality disorders, but none of the other models with the other training sets did. This explains the poor performance of under this model since we only achieved 25% accuracy for class 1 with the synthetic data set. By contrast, the other models placed importance on age, previous hospitalization counts and the total allowed.

Conclusion and Limitations:

The key to the approach for this type of modeling was to notice and address the unbalanced predictand. We can see in the logistic model for the 80/20 split that we would have continued to have very poor performance using the standard methods to achieve classification. Using 4 other different training set (over, under, both, synthetic) we developed a logistic regression, a decision tree, and a random forest.

The logistic regression, the decision tree, and the random forest were all similar in terms of performance. They both had a misclassification rate of about 25%. Random forest while quite similar, distinguishes itself in the ability for it to improve upon the enlargement of the forest size.

The practical implications of our model are that using our classifier, we would be able to predict correctly 75% of the time if a prospective member would be likely to be hospitalized 3 or more times in a year. With this information, the company can then decide the premium and deductible associated with the person's level of risk.

It was mentioned previously that the choice of tuning parameters was very subjective. We chose tuning parameters such that the difference in misclassification errors between classes would be as small as possible, while also choosing the highest general accuracy. We chose this because we wanted to avoid a model that correctly classifies the major class and incorrectly classifies the minor class all the time. We wanted to achieve balance because to achieve a 98% classification rate is not difficult when 98% of the data belongs to the major class. The logistic regression with the 80/20 split achieves this by having a very high accuracy for the major class and a very low accuracy for the minor class.

However, it deserves to be mentioned that there were many good choices for the tuning parameters and we could have chosen one that would favor specificity over sensitivity or vice versa. What we would choose would depend on the focus of the organization. In plainer terms, does the organization prefer to misclassify someone as high risk when they are not? Or do they prefer to misclassify someone as low risk when they are high risk? We suspect that with respect to the bottom line, the company would prefer to have a higher sensitivity than specificity.

With respect to the limitations of the study, we could have increased the tree size in the random forest model. We stopped at 100 because the difference between models was very clear and we did not expect to get a drastic improvement. In addition, the possible range of tuning parameters was very narrow and the number of loops was relatively small (10). We could have created a larger interval with smaller increments and looped over a larger amount of times to get a better estimate for the best choice of parameters.

A further limitation is that there are other types of models that could have been useful here that were not build. We did not build an SVM classifier. In addition, it could be of interest the raw counts of hospitalizations and a multinomial regression or Poisson regression could be useful to those ends.

Lastly, we did not do variable selection. We were interested to see how all of these variables affected the prediction power of the model and therefore did not exclude any to make a more parsimonious model, even upon discovery of statistical insignificance. Had we done this, we could have made marginal improvements to all of our models.

Appendix:

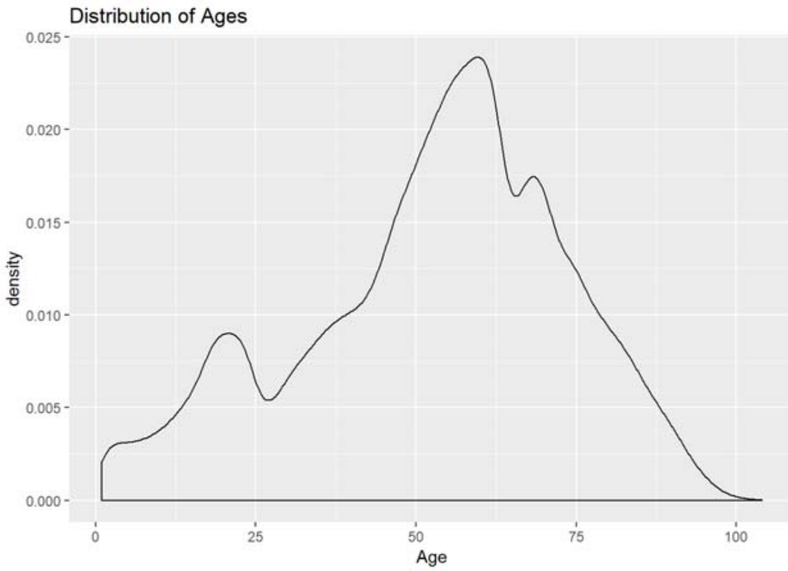
Variable Descriptions:

	Data Type in R	Description	Levels

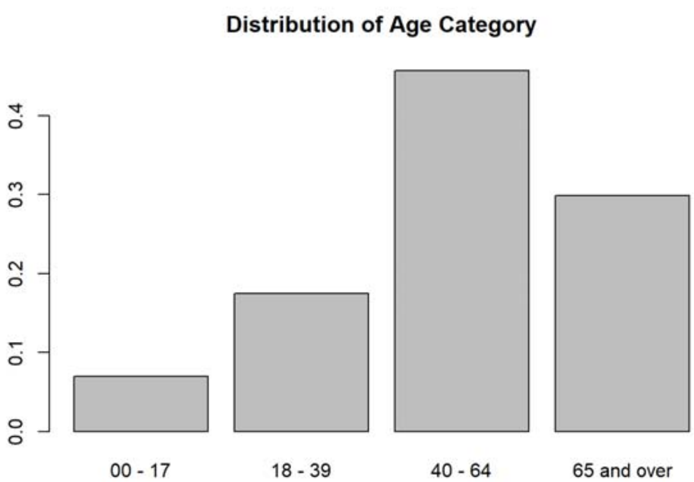
	Data Type in R	Description	Levels
CCMS_ID	numeric	Unique Identification Number	
Age	numeric	Age in 2017	
AgeCat	factor	Age category	00 - 17, 18 - 39, 40 - 64, 65 and over
Gender	factor	Gender	F, M
MEMBER_RELATIONSHIP_DESC	factor	Relationship to Member	child, employee, handicapped dependent, life partner, other relationship, significant other, spouse
Product	factor	Type of insurance Product	commercial, FEP, Med Advantage
Business_Segment	factor	Method of purchase of product	Direct Pay, FEP, Large Group, Med Advantage, Self Insured, Small Group
Hypertension	numeric	Diagnosis of Hypertension	0 = No, 1 = Yes
Hyperlipid	numeric	Diagnosis of Hyperlipid	0 = No, 1 = Yes
LowBackPain	numeric	Diagnosis of lower back pain	0 = No, 1 = Yes
Diabetes	numeric	Diagnosis of diabetes	0 = No, 1 = Yes
IschemicHD	numeric	Diagnosis of Ischemic Heart Disease	0 = No, 1 = Yes
Asthma	numeric	Diagnosis of Asthma	0 = No, 1 = Yes
COPD	numeric	Diagnosis of Chronic Obstructive Pulmonary Disease	0 = No, 1 = Yes
CHF	numeric	Diagnosis of Congestive Heart Failure	0 = No, 1 = Yes
Cancer	numeric	Diagnosis of Cancer	0 = No, 1 = Yes
HIV_AIDS	numeric	Diagnosis of HIV/AIDS	0 = No, 1 = Yes
Depression	numeric	Diagnosis of Depression	0 = No, 1 = Yes
SubstanceAbuse	numeric	Diagnosis of Substance Abuse Disorder	0 = No, 1 = Yes
Schizophenia	numeric	Diagnosis of Schizophrenia	0 = No, 1 = Yes
PersonalityDisorder	numeric	Diagnosis of Personality Disorder	0 = No, 1 = Yes
Bipolar	numeric	Diagnosis of Bipolar Disorder	0 = No, 1 = Yes
Dementia	numeric	Diagnosis of Dementia	0 = No, 1 = Yes
RUB	numeric	Resource Utilization Band	3 = Moderate Morbidity, 4 = High Morbidity, 5 = Very High Morbidity
ACG_Code	numeric	Adjusted Clinical Group Code: Assignation of ICD codes to 32 diagnosis clusters	
PCMH_Program	factor	Patient Center Medical Home: Is the patient in a special coordinated care program?	BCBSRI, CSI, N
Rx_allowed	numeric	Negotiated rate for prescriptions paid out by BCBS	
IP_allowed	numeric	Negotiated rate for inpatient visits	
OP_allowed	numeric	Negotiated rate for outpatient visits	
Prof_allowed	numeric	Negotiated rate for professional visits	
Anc_allowed	numeric	Negotiated rate for ancillary visits (vision, dental)	
IP_Medical_Cnt	numeric	# of medical inpatient visits	
IP_Surgical_Cnt	numeric	# of surgical inpatient visits	

	Data Type in R	Description	Levels
IP_BH_Cnt	numeric	# of inpatient behavioral health visits	
IP_SNF_Cnt	numeric	# of inpatient skilled nursing facility visits	
IP_Rehab_Cnt	numeric	# of inpatient rehab visits	
IP_Maternity_Cnt	numeric	# of inpatient maternity care visits	
IP_Total_Cnt	numeric	Total # of inpatient visits	
IP_Readmit_Cnt	numeric	# of inpatient re-admissions	
Anc_Rad_Cnt	numeric	# of Ancillary radiation visits	
Anc_Lab_Cnt	numeric	# of ancillary lab visits	
OP_ER_Cnt	numeric	# of outpatient ER visits	
OP_Surg_Cnt	numeric	# of outpatient surgical visits	
Prof_UrgCare_Cnt	numeric	# of urgent care center professional visits	
Prof_PCP_Cnt	numeric	# of PCP visits	
Prof_Specialist_Cnt	numeric	# of specialist visits	
Rx_3Mo_Cnt	numeric	# of 3 month prescriptions	
Rx_Specialty_Cnt	numeric	# of specialty prescriptions	
Rx_TotScripts_Cnt	numeric	Total # of prescriptions	
ESRD_Flag	factor	Diagnosis of end stage renal disease	Y, N
Fall_Flag	factor	Significant Risk of Falling	Y, N
Total_Allowed	numeric	Total Amount paid by insurance company under negotiated rate	
Med_Allowed	numeric	Total medical expenses paid by insurance company under negotiated rate	
ihm_npi	factor	Unique provider identifier	
High_Risk_Ind_Std	factor	Risk of Individual	Red, Orange, Other
MRD2	factor	Relationship to policyholder	Dependent, employee
PCMH_Info	factor	Does the patient have a PCP?	Non-PCMH (Has PCP), Non-PCMH (No PCP), PCMH
bccms_id	numeric	Identification #	
OP_ER_Cnt_previous	numeric	Last Years Count of outpatient ER visits	
OP_ER_Cnt_predict	numeric	Predicted count of outpatient ER visits next year	
OP_ER_Cnt_Target	numeric	Was the patient hospitalized >3 times this year	0 = Less than 3 outpatient hospitalizations, 1 = Three or greater outpatient visits.

Distributions of Variables:

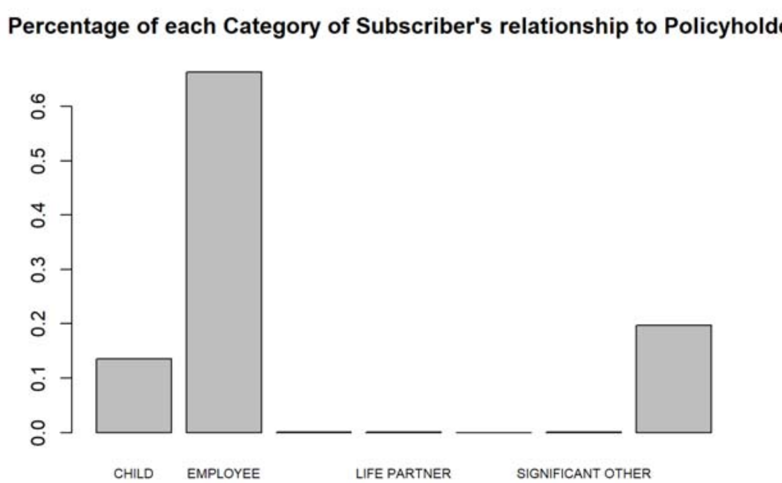


Comments: This has a slight left tail, but it is not too extreme.

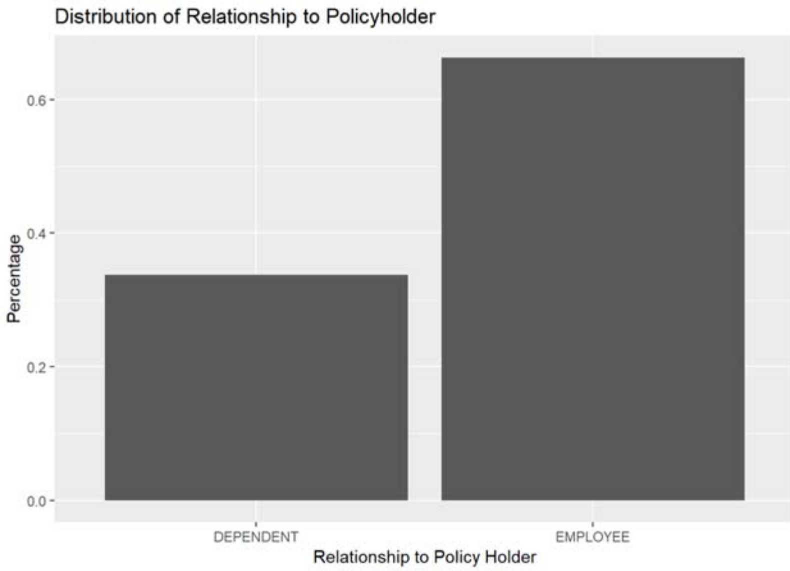


Comments: This reflects the same as above. People under 40 compose only ~25% of the data.

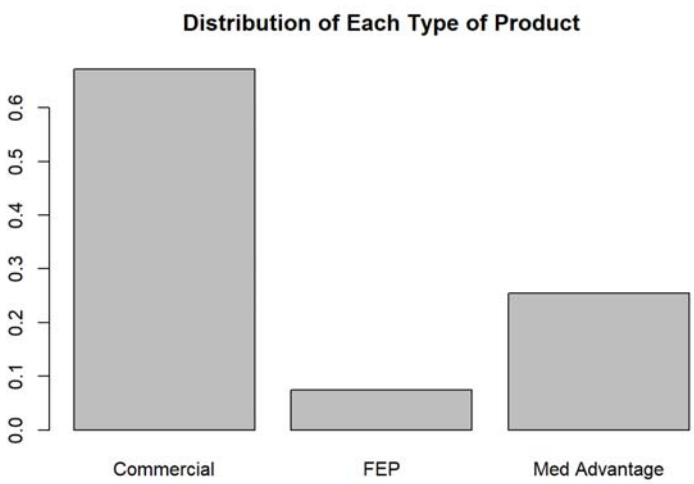
Comments: The gender data is close to being balanced with women composing 57% of the data and men 43%.



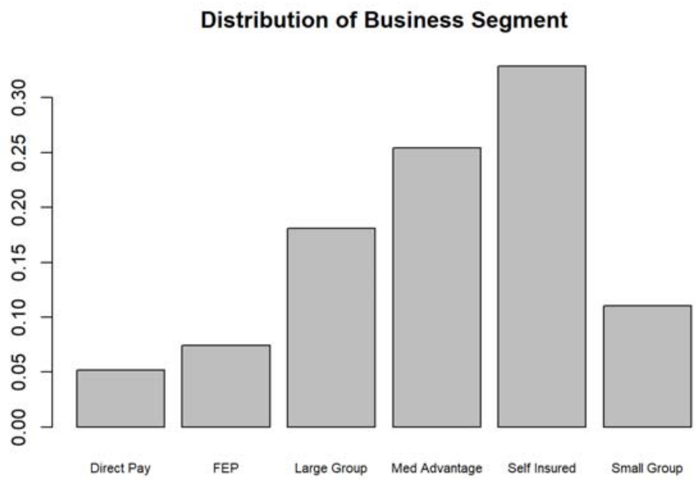
Comments: Originally there were only 4 members in the “Other Relationship” category. There were removed because they could not be placed in any other categories and they were too rare to be guaranteed to be in both training and testing sets.



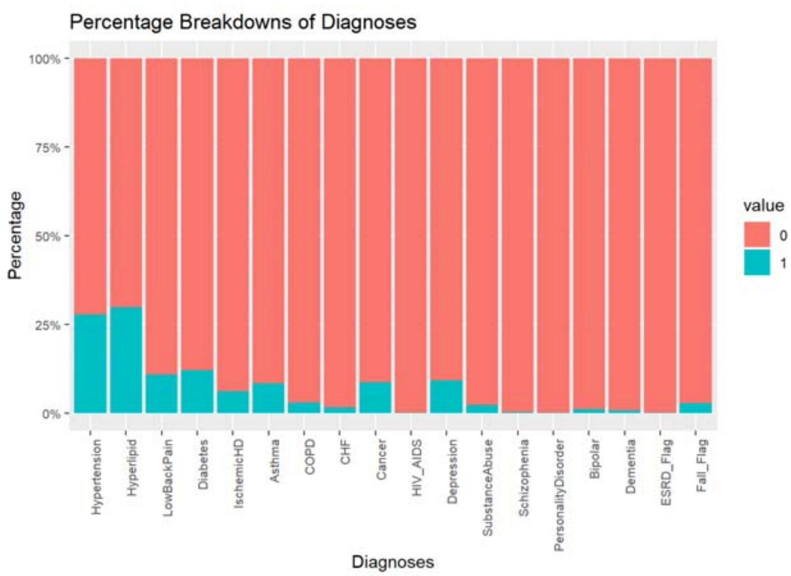
Comments: Most people in the data set are the policy holder. About 1/3 are dependents of the policyholder.



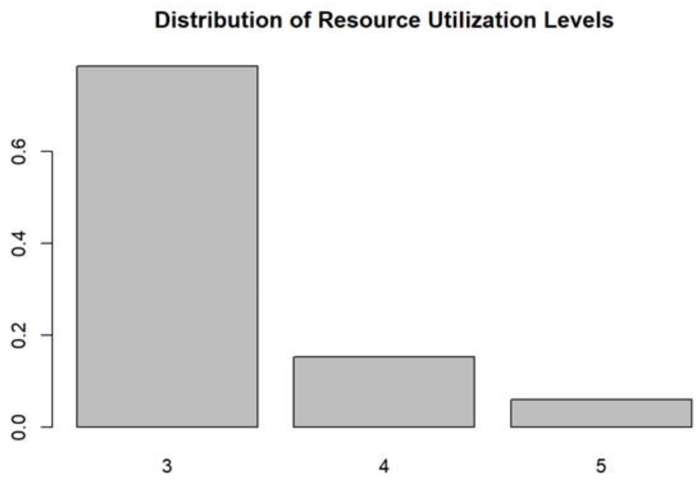
Comments: Most members have insurance purchased commercially and not through their government job or through Medicare.



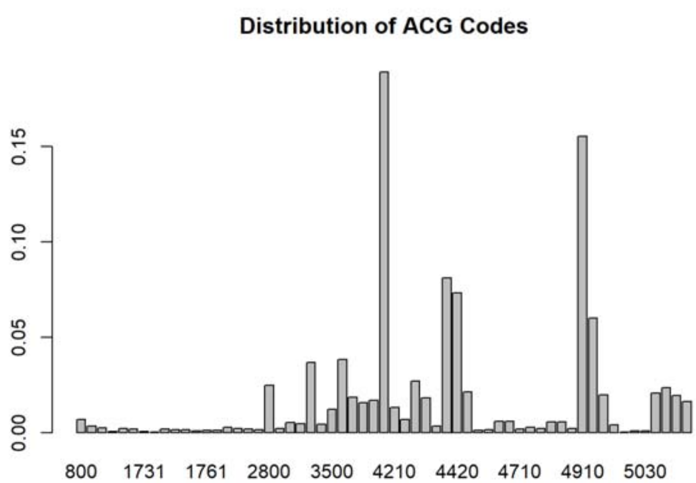
Comments: The majority of people in the data set have medicare or receive insurance through their place of employment. Very few buy their own plans (Direct Pay) or have insurance through their government jobs. (FEP)



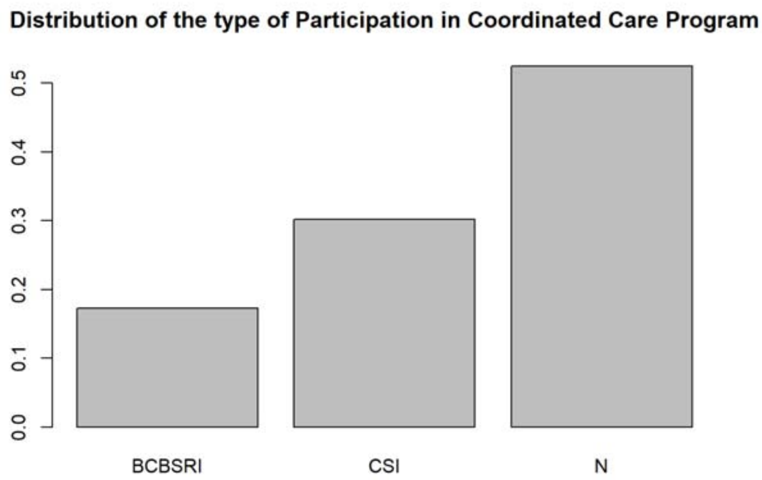
Comments: The most common diagnoses are hypertension and hyper-lipids with about 25% of the population in the data set having those diagnosis.



Comments: Most members are categorized as having a moderate morbidity. Relatively very few other members are categorized at being higher risk.

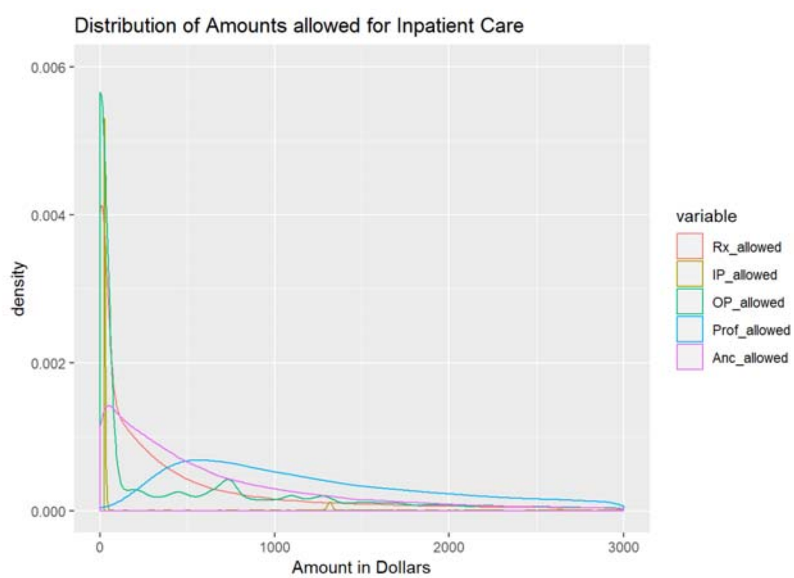


Comments: The most frequent ACG codes are 4100, 4910, 4410. These codes all correspond to people who have no major illnesses, which aligns well the rest of the data available.

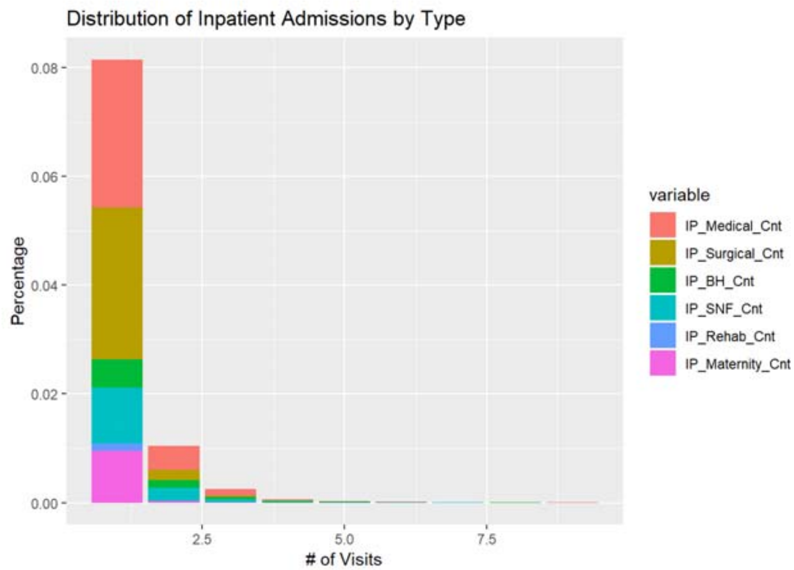


Comments: Most people do not see PCP's (or it is not known if they see PCP's) who are a part of a special coordinated care program.

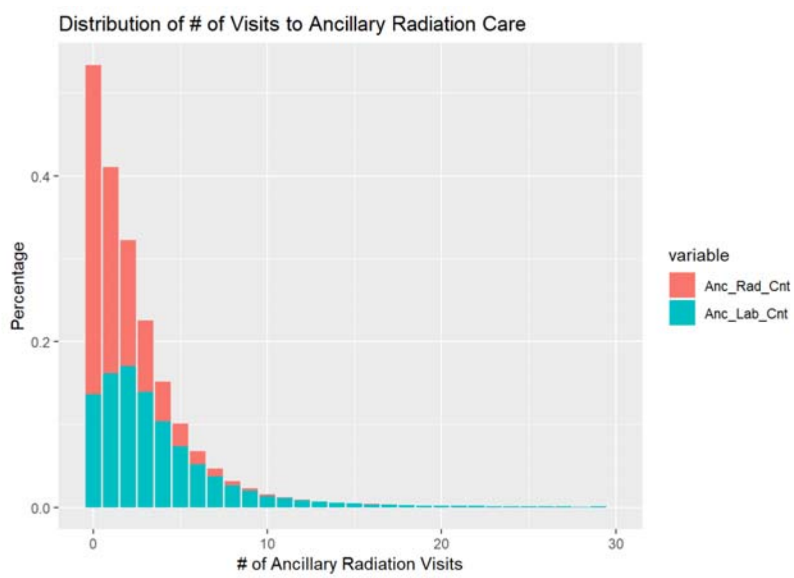
```
## Warning: Removed 87747 rows containing non-finite values (stat_density).
```



Comments: All of these distribution are extremely skewed with a heavy right tail. The distributions plotted here include only the data for for dollar amounts less than $3000 since the tails extended well past $1,000,000. The warning listed above is because only amounts less than $3000 were considered for graphical purposes and the rest was thrown away. ~10% of the data was ignored.



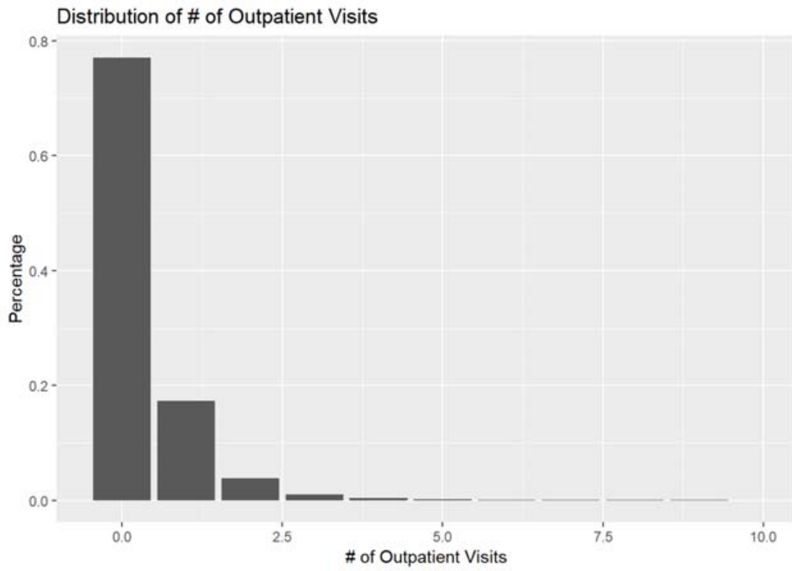
Comments: There is still a heavy right tail in this distribution. The bulk of inpatient admissions are for surgical or medical reasons.



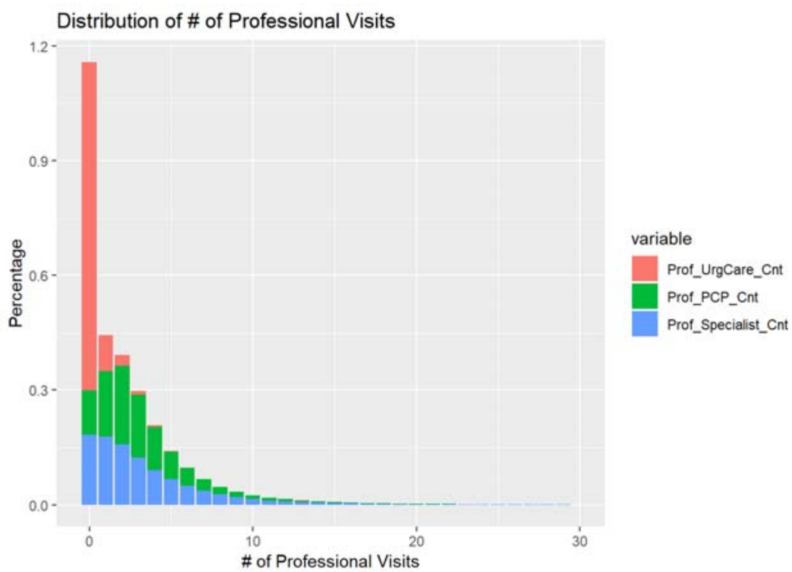
Comments: Right tailed. Heavily skewed. The majority of Ancillary visits are for labs.

```
## Warning: Removed 98 rows containing non-finite values (stat_count).
```

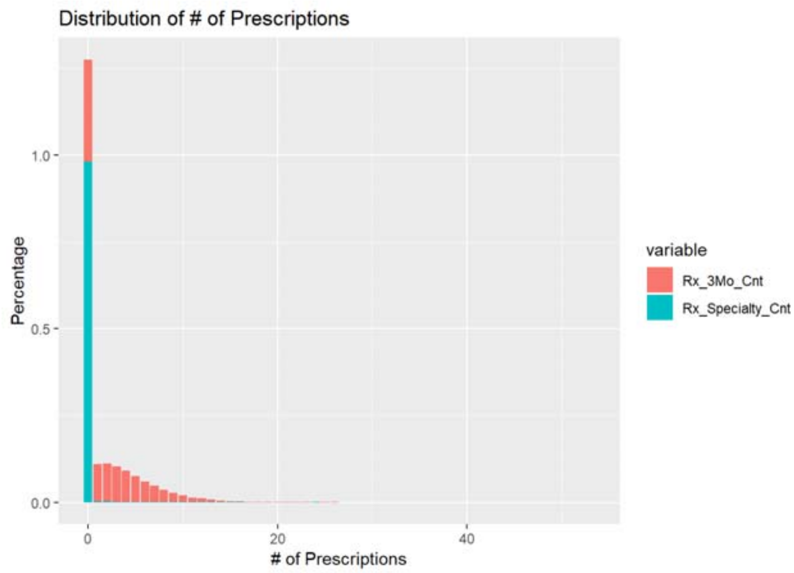
```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



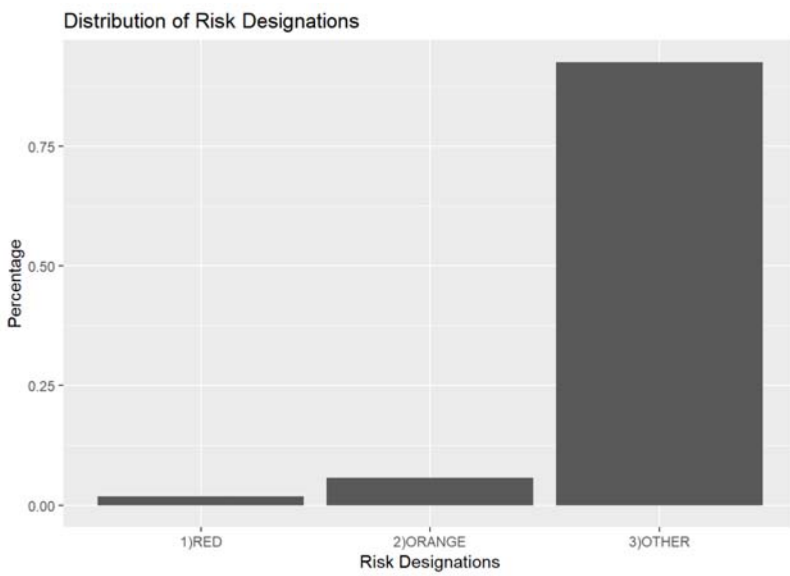
Comments: Right tailed. Heavily skewed. It's a 50/50 split between the ER and surgical visits.



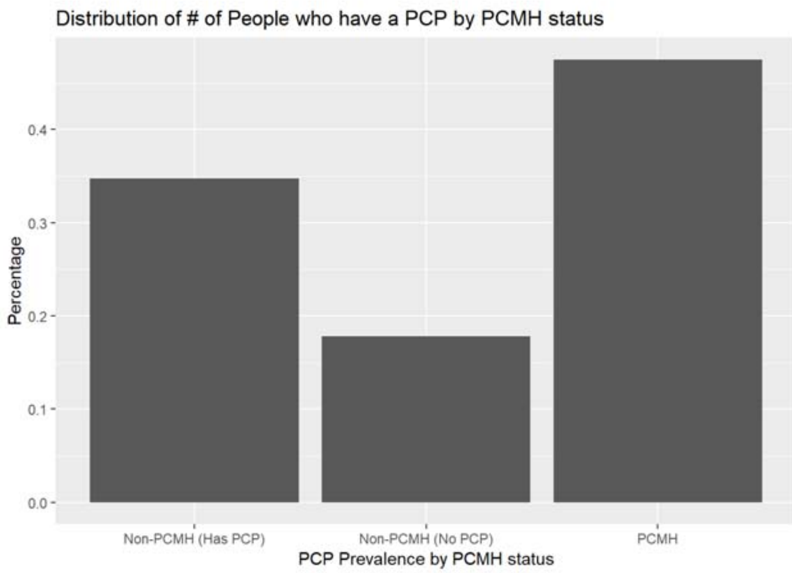
Comments: More right skewed than some of the other plots. Most professional visits are either to a PCP or a specialist. Urgent care is quite rare.



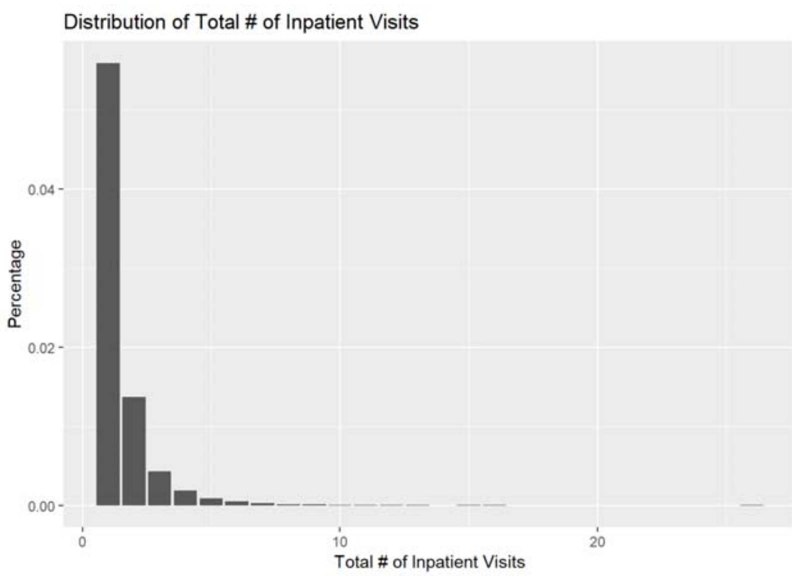
Comments: Right skewed and most are regular 3 month prescriptions.



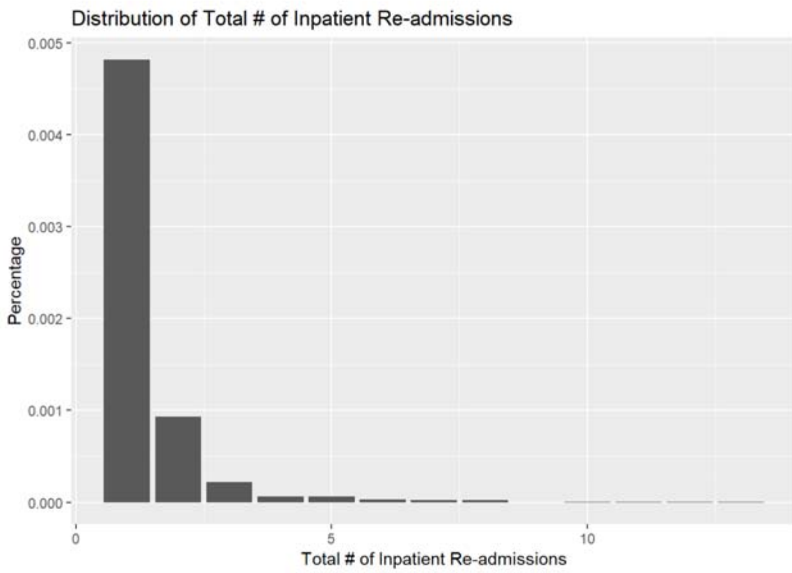
Comments: I have no idea what the “Other” category means.



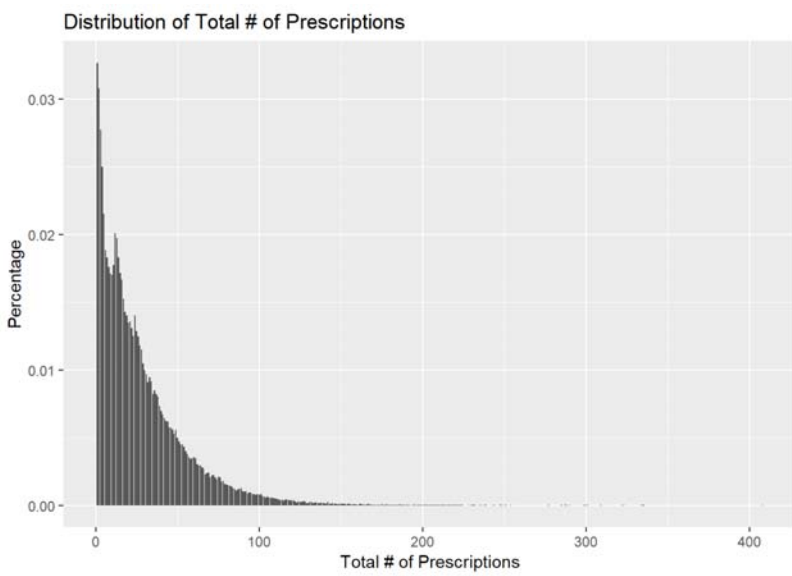
Comments: About 75% have a PCP regardless of PCMH status. Everyone in a PCMH program has a PCP. About 2/3 have a PCP for those *not* in a PCMH program.



Comments: Right tailed. Heavily skewed distribution.This plot only shows >0 inpatient visits for graphical clarity, which is why it does not sum to 1. The distribution is even more right tailed when looking at *all* the possible counts.

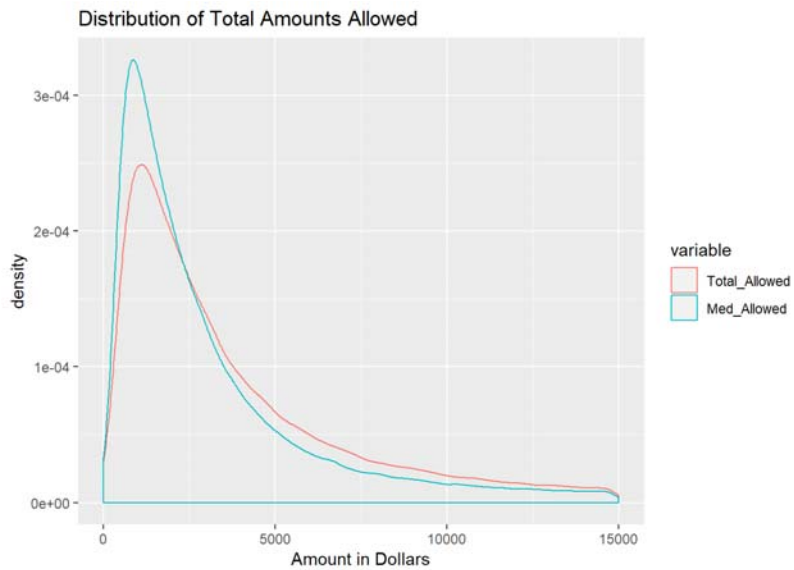


Comments: Right tailed. Heavily skewed distribution.This plot only shows >0 inpatient visits for graphical clarity. The distribution is even more right tailed when looking at *all* the possible counts.



Comments: Heavy right tail. This graph only shows values greater than 0. Therefore, the full distribution is even *more* right tailed.

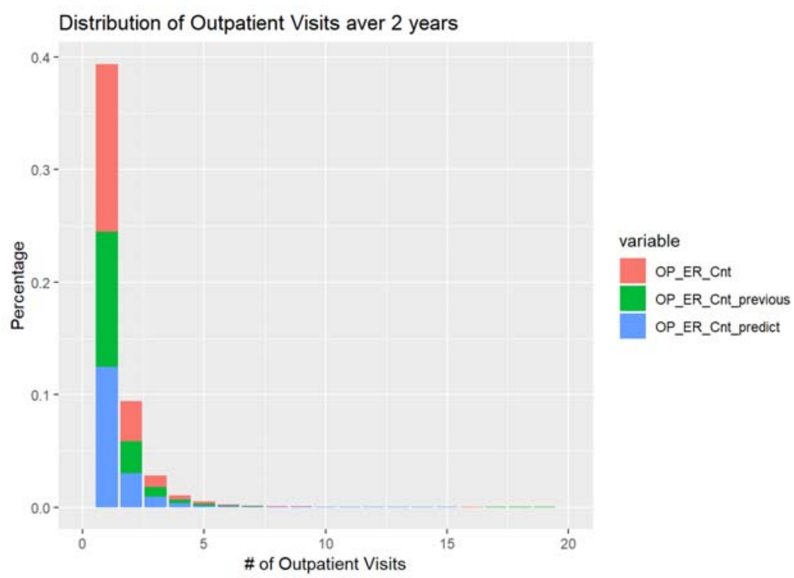
```
## Warning: Removed 35869 rows containing non-finite values (stat_density).
```



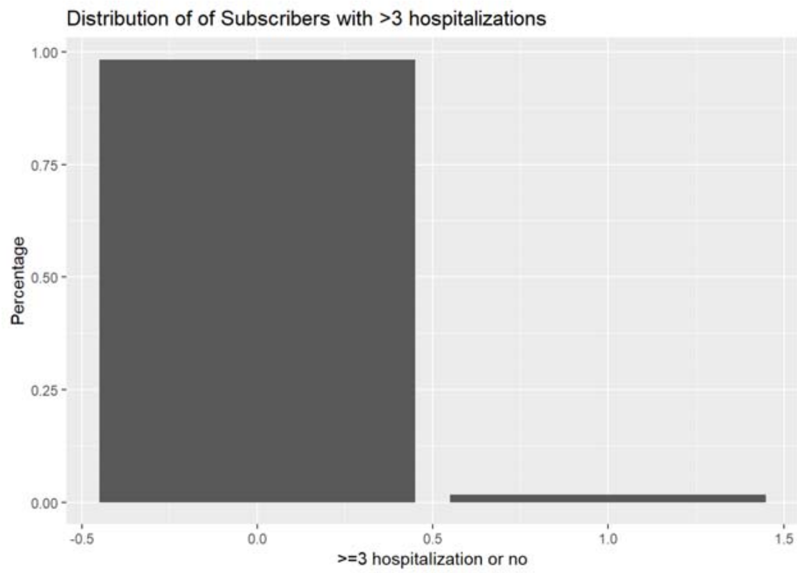
Comments: Heavy right tail as expected.

```
## Warning: Removed 36 rows containing non-finite values (stat_count).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Comments: This distribution is also right skewed. We can see that amount of outpatient visits is the same between years, and it is also consistent with the predicted amount.



Comments: This is a very unbalanced data set. The majority of people do *not* have 3 or more hospitalizations.