Paper 1031-2017

Analyzing the Effectiveness of COPD Drugs Through Statistical Tests and Sentiment Analysis

Indra kiran Chowdavarapu, Oklahoma State University; Dr. Scott Shepherd, OSU CHSI;
Dr. Miriam McGaugh, Oklahoma State University

ABSTRACT

Chronic Obstructive Pulmonary disease(COPD) is the third leading causing of death in the United States. An estimated 24 million people suffer from COPD, and the medical costs associated with it stands at a whopping $36 billion. Besides the emotional and physical impact, a patient with COPD has to undergo severe economic burden to pay for the medication. At this juncture, identifying the best medicine to treat COPD enhances the living conditions of patients. This paper deals with analyzing the effectiveness of three popular drugs prescribed for the COPD patients through statistical tests and Sentiment analysis. The statistical analysis determines the effectiveness of these drugs on the patients in terms of mortality rates and readmission within 30 days of discharge. The impact of comorbidities, such as cardio vascular diseases, accident history, and smoking, on COPD patients is also examined. The data consists of 1 million patient encounter records obtained through Cerner Health Facts data. Base SAS is used to perform statistical analysis, combine multiple data sets to obtain each patient's hospital records, compute readmission within 30 days' information using lag function, and perform descriptive analysis. This paper also includes text mining of patients' reviews about the drugs on drug portals and social media. The results obtained through sentiment analysis are then compared with the results of statistical analysis obtained earlier to determine the effectiveness of drugs prescribed to the COPD patients.

INTRODUCTION

Chronic Obstructive Pulmonary Disease, is a chronic lung disease that makes it hard to breathe because less air flows in and out of the airways in your lungs. When you're getting less air, less oxygen gets into body tissues and it gets harder to get rid of the waste gas carbon dioxide. This results in shortness of breath during everyday activities. People with COPD can experience fatigue, chronic cough and frequent respiratory infections as well. COPD doesn't just have a physical impact—living with chronic disease also can affect mental health. The current research is aimed at determining three objectives:

1. Analyze the effectiveness of 3 COPD drugs: Advair, Symbicort, Spiriva, the combination of Advair/Spiriva and Symbicort/Spiriva. The effectiveness of these drugs are tested in terms of patients

   i. Length of Stay in the hospital

   ii. Readmission within 30 days of discharge

   iii. Mortality from COPD

2. Understand the effect of various Comorbidities on the outcome of COPD patients. The Comorbidities tested in this research are Accidents, Heart diseases and Smoking. The influence of these Comorbidities are seen in terms of COPD patients

   i. Length of Stay in the hospital

   ii. Readmission within 30 days of discharge

   iii. Mortality from COPD and Comorbidity

3. Sentiment analysis on drug portals to determine the sentiments of patients using Advair, Symbicort and Spiriva.

DATA PREPARATION

The Cerner Health Facts data set used in this project is a comprehensive collection of Patient information, Discharge information, Hospital information, Medication information of patients diagnosed with COPD. The ICD codes used for the analysis are given below:

DIAGNOSIS_CODE	DIAGNOSIS_DESCRIPTION
490	Obstructive Chronic Bronchitis with (Acute) Exacerbation
491	Other Emphysema
491.1	Other Emphysema
491.2	Asthma, Unspecified
491.21	Asthma, Unspecified
491.22	Chronic Airway Obstruction, Not Elsewhere Classified
491.8	Obstructive Chronic Bronchitis with (Acute) Exacerbation
491.9	Other Emphysema
492	Obstructive Chronic Bronchitis with (Acute) Exacerbation
492.8	Asthma, Unspecified

Table 1. Diagnosis codes used in the analysis

The date set consists of a total of 593,932 patient records. These records include patients with all the medications prescribed for the patients admitted with COPD. For the current scope of study, we have identified the patients who have been given either of Advair, Symbicort, Spiriva or combination of Advair/Symbicort with Spiriva. Some of the patients were prescribed these medications for other ailments other than COPD. To eliminate this problem, the patient records are extracted based on the above ICD codes. The final data set consists of 83,395 unique patient records have used the mentioned drugs with diagnostic codes related to analysis. Figure 2 represents the selection process of the data for the purpose of this study.

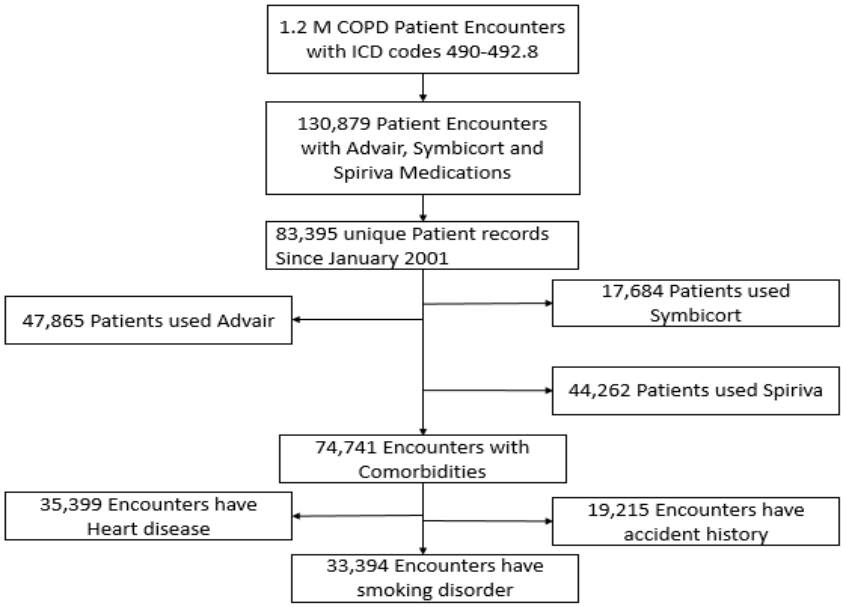


Figure 2. Overview of data used in the study

TARGET VARIABLES

For the current analysis, three target variables are created from the existing data set

1. Length of stay of patient:

The variable length of stay is calculated by subtracting the discharge date of patient with his admitted time. The length of stay is calculated at the level of patient's each encounter to the hospital but not on overall visits of patient to the hospital.

2. Mortality flag for Mortality analysis:

The variable Mortality_Flag is extracted from the column Discharge_disposition_id in the date set. The Discharge_disposition_id is a categorical variable that contains integer numbers for each of the discharge types. For eg, the discharge to Skilled Nursing facility is denoted by integer 12. In the current analysis, the main focus is on the final status of the patient irrespective of the discharge facility patient has been prescribed. In order to achieve that, all the patients who are discharged from the hospital are grouped as Target=0. The patients who were expired during or after the treatment are categorized into Target=1. Some of the levels of target variables can be seen below.

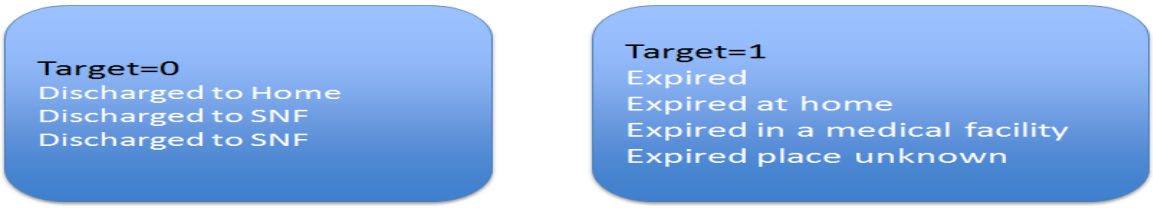


Figure 2. Description of Mortality_flag variable

3. Readmission flag for readmission analysis:

The variable readmit30 is created from the existing data set. The variable is a binary variable that defines the readmission status of the patient with COPD.

Readmit30='0' => The patient is not readmitted into the hospital within 30 days of discharge from the hospital
Readmit30='1' => The patient is readmitted into the hospital within 30 days of discharge from the hospital

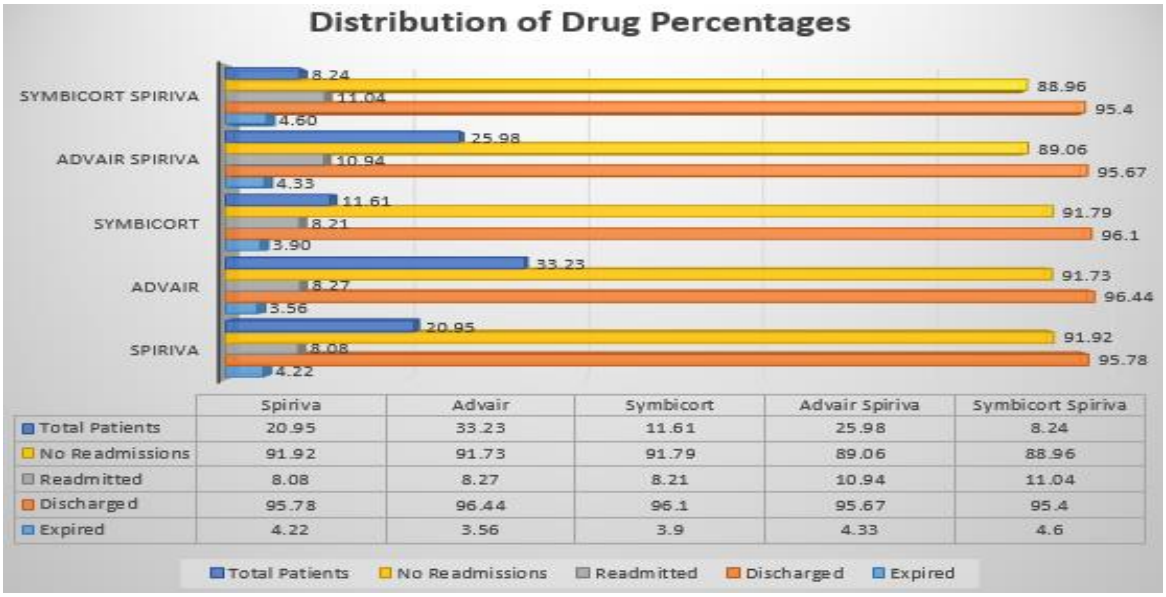
The below analysis is used to identify patients who are readmitted to hospital within 30 days of discharge from the hospital. The steps followed for the analysis is shown below:

```
/*Calculating Length of Stay and converting date format to calculate
readmission*/
data &out ;
set &in ;
      date_dis = datepart(discharged_dt_tm);
      /*format date mmddyy10.;*/
      date_adm = datepart(admitted_dt_tm);
      /*format date mmddyy10.;*/
      length_stay = date_dis - date_adm;
run;
```

```
/* Identifying patients with readmission within 30 days of discharge*/
data copd.indra_rank;
set &out;
Rank = 1;
run;

PROC SORT DATA=copd.indra_rank;
  BY patient_sk DESCENDING date_adm date_dis;
RUN ;

data copd.indra_seq;
  set copd.indra_rank;
  by patient_sk;
  if first.patient_sk then Seq_No=0;
Seq_No+rank;
run;

PROC SORT DATA=copd.indra_seq OUT=copd.indra_seq ;
  BY patient_sk Seq_No;
RUN ;

data copd.indra_seq;
set copd.indra_seq;
by patient_sk Seq_No;
ref_dt_id=LAG(date_adm);
gap = ref_dt_id - date_dis;
length_of_stay = date_dis - date_adm;
if First.patient_sk then do;
     ref_dt_id =.;
     Gap =.;
     readmit30 =.;
End;
If 0 <= Gap <= 30 then
readmit30 = 1;* Identify a readmission and assign value 1 to readmit30;
else readmit30 = 0;
Run;
```

STATISTICAL TESTS TO MEASURE EFFECTIVENESS OF DRUGS

The drugs used in the current study are Advair, Symbicort, Spiriva and combination of Advair/Symbicort with Spiriva. To eliminate the biases, each combination of drug is considered individually. Patient who has taken combination of Advair and Spiriva is considered different from patient who has taken only Advair or Symbicort. The distribution of medications on various measures can be seen in the figure 2.



	Spiriva	Advair	Symbicort	Advair Spiriva	Symbicort Spiriva
Total Patients	20.95	33.23	11.61	25.98	8.24
No Readmissions	91.92	91.73	91.79	89.06	88.96
Readmitted	8.08	8.27	8.21	10.94	11.04
Discharged	95.78	96.44	96.1	95.67	95.4
Expired	4.22	3.56	3.9	4.33	4.6

Figure 2. Distribution of drug percentages among the patients

1. LENGTH OF STAY:

The length of stay is computed for each patient's encounter to the hospital. ANOVA test is performed to understand the effectiveness of drugs on patient's length of stay. The ANOVA test compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

H0: μ1= μ2= μ3….= μk

where μ = group mean and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (HA), which is that there are at least two group means that are statistically significantly different from each other.

```
/*ANOVA test to determine effectiveness of medicines*/
proc anova data=&data;
class medication;
model length_of_stay=medication;
means medication/ duncan;
run;
```

The ANOVA Procedure

Dependent Variable: length_of_stay

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	10778.754	2694.688	59.11	<.0001
Error	129999	5926335.104	45.588		
Corrected Total	130003	5937113.857			

Figure 3. ANOVA procedure on length of stay

The p-value for the ANOVA is less than 0.05 which means at least two groups have means different from others.

Post-hoc Tests:

The one-way ANOVA is an omnibus test statistic and cannot tell you which specific groups were statistically significantly different from each other, only that at least two groups were. To determine which specific groups differed from each other, you need to use a post hoc test. For this analysis, Duncan's test is used to determine the differences between the groups.

This Post-hoc test determines if there exists difference between certain groups of Medication. As seen from the output, there is no statistically significant difference for Spiriva and Symbicort. The combination of Symbicort and Spiriva, Advair and Spiriva and Advair are significantly different from each other. The medications Symbicort combined with Spiriva has the highest length of stay for a patient with an average of 6.54 days per visit.

Out of all the medications, Advair has the best length of stay with an average of 5.72 days per visit.

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	medication
A	6.54540	10706	symbicort spiriva
B	6.36992	33775	advair spiriva
C	6.00698	27236	spiriva
C			
C	5.97906	15090	symbicort
D	5.72480	43197	advair

Figure 4. Output for Duncan's Post hoc test

Assumptions of ANOVA:

To use ANOVA results, we need to test if it is satisfying various assumptions. One of the assumption tested in this analysis is to verify if there is any heteroscedasticity in the data.

```
/*Testing and adjusting for unequal variances (heteroscedasticity)*/
proc anova data=&data;
class medication;
   model length_of_stay = medication;
   means medication / hovtest welch;
run;
```

Levene's Test for Homogeneity of length_of_stay Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
medication	4	11087146	2771787	2.68	0.0300
Error	82813	8.569E10	1034786		

Welch's ANOVA for length_of_stay			
Source	DF	F Value	Pr > F
medication	4.0000	52.18	<.0001
Error	27400.2		

Figure 5. Levene's test for homogeneity of variance

The P-value for the tests are less than .05. Hence, the assumption of equal variance is sustained.

2. MORTALITY ANALYSIS:

Total Number of Patients considered: 83394 (130986 encounters)

No. of patients discharged: 80044 (96.7%)

No. of patients expired: 3350 (3.3%)

Chi-square test for the relationship between mortality and medications:

The Chi-square test is used to test if any relationship exists between two categorical variables. In our current analysis, Chi-square test can be used to understand if type of drug prescribed for the patient has any outcome in improving patient's chances of recovery from COPD.

Assumptions:

Null Hypothesis: There is no association between Drug administered and mortality of the patient.

Alternate Hypothesis: There is an association between Drug administered and mortality of the patient.

```
/*chi-sq tests on Mortality*/
proc freq data=&data;
tables mortality_flag * medication/chisq norow nopercent;
run;
```

Frequency Col Pct	Table of mortality_flag by medication					
	medication					
mortality_flag	advair	advair spiriva	spiriva	symbicort	symbicort spiriva	Total
0	27179 96.44	18286 95.67	17634 95.78	10261 96.10	6140 95.40	79500
1	1002 3.56	827 4.33	777 4.22	416 3.90	296 4.60	3318
Total	28181	19113	18411	10677	6436	82818

Statistics for Table of mortality_flag by medication

Statistic	DF	Value	Prob
Chi-Square	4	28.4037	<.0001
Likelihood Ratio Chi-Square	4	28.5534	<.0001
Mantel-Haenszel Chi-Square	1	12.6902	0.0004
Phi Coefficient		0.0185	
Contingency Coefficient		0.0185	
Cramer's V		0.0185	

Sample Size = 82818

Figure 6. Chi-square test on Mortality_flag and Medication

The chi-square value is 0.001 which is less than the p-value of 0.05. It means with 95% confidence we can say that, there is an association between the type of drug administered to the patient and the chances of survival of patient.

Multiple Comparison tests on proportions:

Tukey's multiple comparison test is performed to determine if there exist any differences between different groups of medicines on mortality.

Tukey Style Multiple Comparisons of Proportions

Compare	Diff	SE	q	q(.05)	Conclude
1 vs 5	1.52	0.28	5.44	3.585	R
1 vs 2	1.14	0.19	5.99	3.585	R
1 vs 3	0.99	0.19	5.2	3.585	R
1 vs 4	0.52	0.23	2.26	3.585	A
4 vs 5	1	0.32	3.13	3.585	A
4 vs 2	0.62	0.24	2.58	3.585	A
4 vs 3	0.47	0.25	1.87	3.585	A
3 vs 5	0.53	0.29	1.84	3.585	A
3 vs 2	0.15	0.21	0.72	3.585	A
2 vs 5	0.38	0.29	1.32	3.585	A

Figure 7. Output for Tukey's Post hoc test

Out of the various medications, the differences in proportions for 1 vs 5 (Advair vs Symbicort Spiriva), 1 vs 2 (Advair vs Advair Spiriva) and Advair vs Spiriva are significant.

ODDS RATIO:

An odds ratio (OR) is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

```
/*ODDS Ratio for Mortality Analysis*/
Proc logistic data=indra.medicine_mortal_final;
class medication(ref = 'advair') ;
model mortality_flag(event='1') = medication;
run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
medication advair spiriva vs advair	1.227	1.117	1.348
medication spiriva vs advair	1.195	1.086	1.315
medication symbicort vs advair	1.100	0.979	1.236
medication symbicort spiriva vs advair	1.308	1.145	1.493

Figure 8. Odds ratio of Medications in determining Mortality

The Odds of Mortality ('1') is 22% more for combination of Advair, Spiriva than for Advair alone. Based on all the odds, Advair has the less outcome for Mortality.

Male vs Female:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
GENDER Male vs Female	1.253	1.169	1.343

Figure 9. Odds ratio of Gender in determining Mortality

Urban vs Rural:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
URBAN_RURAL_STATUS Rural vs Urban	1.138	1.041	1.244

Figure 10. Odds ratio of location in determining Mortality

3. READMISSION ANALYSIS:

Total Number of Encounters considered: 130004

No. of patients discharged: 118112 (90.85%)

No. of patients readmitted within 30 days: 11892 (9.14%)

A total of 11892 patients were readmitted within 30 days of discharge from hospital. This comprises of 9.14% of total patients admitted into hospital with COPD

Chi-square test for the relationship between two categorical variables:

Frequency Col Pct	Table of readmit_flag by medication					
	medication					
readmit_flag	advair	advair spiriva	spiriva	symbicort	symbicort spiriva	Total
0	39623 91.73	30079 89.06	25035 91.92	13851 91.79	9524 88.96	118112
1	3574 8.27	3696 10.94	2201 8.08	1239 8.21	1182 11.04	11892
Total	43197	33775	27236	15090	10706	130004

Statistics for Table of readmit_flag by medication

Statistic	DF	Value	Prob
Chi-Square	4	270.0623	<.0001
Likelihood Ratio Chi-Square	4	263.5347	<.0001
Mantel-Haenszel Chi-Square	1	9.7141	0.0018
Phi Coefficient		0.0456	
Contingency Coefficient		0.0455	
Cramer's V		0.0456	

Sample Size = 130004

Figure 11. Chi-square test on Readmit within 30 days and Medication

The chi-square value is 0.001 which is less than the p-value of 0.05. It means with 95% confidence we can say that, there is an association between the type of drug administered to the patient and the chances of readmission of patient within 30 days of discharge.

Post Hoc Test:

```
Tukey Style Multiple Comparisons of Proportions

Compare Diff SE q q(.05) Conclude
------------------------------------------------------------
   3 vs 5    2.89      0.23      12.58     3.585     R
   3 vs 2    2.8       0.16      17.51     3.585     R
   3 vs 1    0.2       0.16      1.25      3.585     A
   3 vs 4    0.14      0.21      0.65      3.585     A
   4 vs 5    2.76      0.26      10.6      3.585     R
   4 vs 2    2.66      0.2       13.32     3.585     R
   4 vs 1    0.06      0.19      0.34      3.585     A
   1 vs 5    2.69      0.22      12.24     3.585     R
   1 vs 2    2.6       0.15      17.34     3.585     R
   2 vs 5    0.09      0.22      0.42      3.585     A
```

Figure 12. Output for Tukey's Post hoc test

The following post hoc test determines the three drugs Spiriva, Spiriva Symbicort and Advair to be Statistically significant.

ODDS RATIO:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
medication advair spiriva vs advair	1.362	1.298	1.430
medication spiriva vs advair	0.975	0.922	1.030
medication symbicort vs advair	0.992	0.927	1.061
medication symbicort spiriva vs advair	1.376	1.284	1.475

Figure 13. Odds ratio of medications in determining readmission

The Odds ratio determines Spiriva to have the least odds for the readmission.

Male vs Female:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
GENDER Male vs Female	1.253	1.169	1.343

Figure 14. Odds ratio of gender in determining readmission

Urban vs Rural:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
URBAN_RURAL_STATUS Rural vs Urban	1.138	1.041	1.244

Figure 15. Odds ratio of location in determining readmission

II. STATISTICAL TESTS TO MEASURE IMPACT OF COMORBIDITIES

The Comorbidities tested in this research are Accident_history, Heart diseases and Smoking disorder. The influence of these Comorbidities are measured in terms of length of stay, readmission within 30 days and mortality of COPD patients.

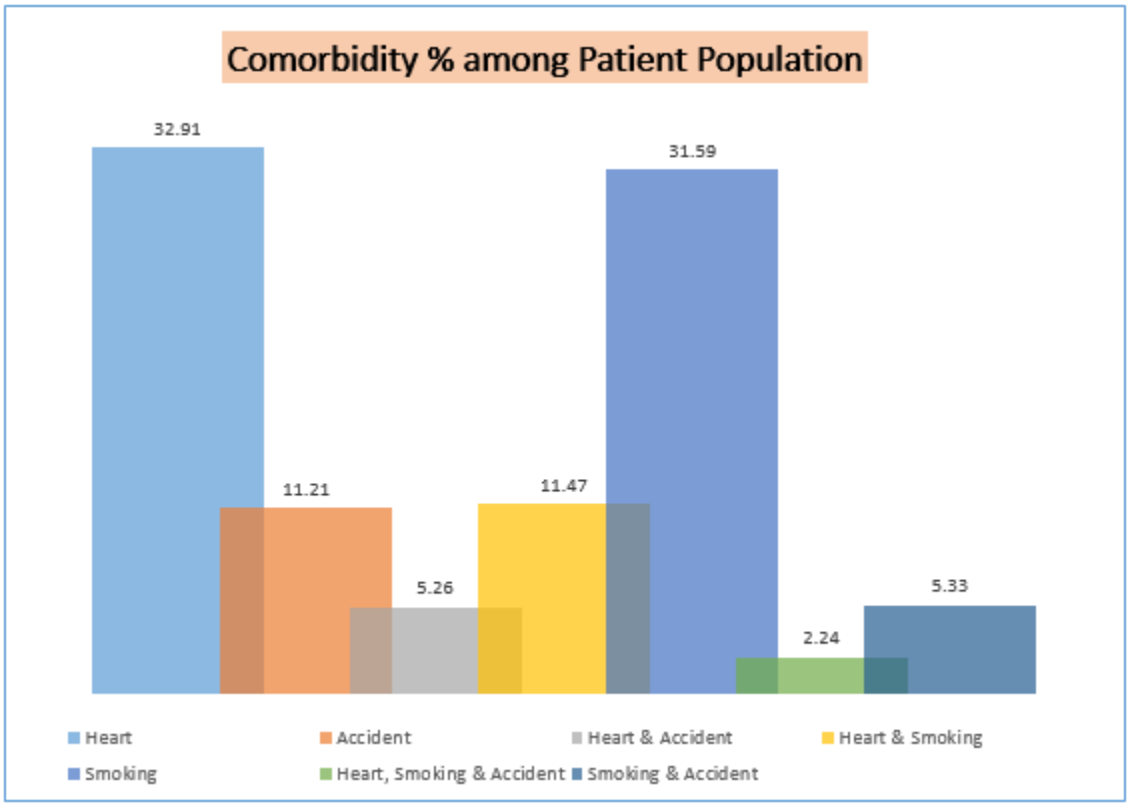


Figure 16. Percentage of Comorbidities among COPD patient population

1. LENGTH OF STAY:

ANOVA test is performed to understand the impact of comorbidities on patient's length of stay.

The ANOVA Procedure

Dependent Variable: length_of_stay

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	116669.127	19444.854	431.52	<.0001
Error	74734	3367584.719	45.061		
Corrected Total	74740	3484253.846			

Figure 17. ANOVA test on comorbidities in determining length of stay

The p-value for the ANOVA is less than 0.05 which means at least two groups have means different from others.

Post-hoc Tests:

The ANOVA Procedure

Duncan's Multiple Range Test for length_of_stay

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	complication
A	8.9443	3969	Heart Accident
A			
A	8.8032	1677	Heart Smoking Accident
B	7.9006	8412	Accident
C	7.2244	3962	Smoking Accident
D	5.8756	24720	Heart
E	5.3227	8558	Heart Smoking
F	4.8968	23443	Smoking

Figure 18. Output of Duncan's multiple range test

The Comorbidities are significantly different from each other. The Comorbidity with Accident_history has the highest length of stay and Smoking disorder has the least length of the stay.

2. EFFECT OF COMORBIDITIES ON MORTALITY:

Chi-square test for the relationship between mortality and comorbidities:

Frequency
Col Pct

Table of mortality_flag by complication

mortality_flag	complication							
	Accident	Heart	Heart Accident	Heart Smoking	Heart Smoking Accident	Smoking	Smoking Accident	Total
0	7922 96.68	23381 97.17	3647 94.88	8273 98.62	1577 96.28	22796 98.71	3821 98.07	71417
1	272 3.32	681 2.83	197 5.12	116 1.38	61 3.72	299 1.29	75 1.93	1701
Total	8194	24062	3844	8389	1638	23095	3896	73118

Statistic	DF	Value	Prob
Chi-Square	6	352.8403	<.0001
Likelihood Ratio Chi-Square	6	337.6340	<.0001
Mantel-Haenszel Chi-Square	1	176.4192	<.0001
Phi Coefficient		0.0695	
Contingency Coefficient		0.0693	
Cramer's V		0.0695	

Effective Sample Size = 73118

Figure 19. Chi-square test on mortality_flag and comorbidities

The Chi-square value is significant. There is association between mortality and Comorbidity.

Tukey's Post-hoc Tests:

Tukey Style Multiple Comparisons of Proportions

Compare	Diff	SE	q	q(.05)	Conclude
6 vs 3	6.56	0.35	18.75	4.17	R
6 vs 5	4.63	0.52	8.91	4.17	R
6 vs 1	3.97	0.26	15.26	4.17	R
6 vs 2	3.15	0.19	16.58	4.17	R
6 vs 7	1.46	0.35	4.18	4.17	R
6 vs 4	0.23	0.26	0.88	4.17	A
4 vs 3	6.33	0.39	16.24	4.17	R
4 vs 5	4.4	0.55	8	4.17	R
4 vs 1	3.74	0.31	12.07	4.17	R
4 vs 2	2.92	0.26	11.24	4.17	R
4 vs 7	1.23	0.39	3.16	4.17	A
7 vs 3	5.1	0.46	11.09	4.17	R
7 vs 5	3.17	0.6	5.28	4.17	R
7 vs 1	2.51	0.39	6.43	4.17	R
7 vs 2	1.69	0.35	4.82	4.17	R
2 vs 3	3.41	0.35	9.75	4.17	R
2 vs 5	1.48	0.52	2.85	4.17	A
2 vs 1	0.82	0.26	3.15	4.17	A
1 vs 3	2.59	0.4	6.48	4.17	R
1 vs 5	0.66	0.55	1.2	4.17	A
5 vs 3	1.93	0.6	3.22	4.17	A

Figure 20. Output for Tukey's Post hoc test

Based on the Tukey's test on multiple proportions, the Comorbidities Smoking, Heart-Smoking, Smoking-Accident and Accident have significant association.

ODDS RATIO:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
complication Accident vs Heart	1.179	1.022	1.360
complication Heart Accident vs Heart	1.855	1.577	2.182
complication Heart Smoking vs Heart	0.481	0.395	0.587
complication Heart Smoking Accident vs Heart	1.328	1.017	1.734
complication Smoking vs Heart	0.450	0.393	0.517
complication Smoking Accident vs Heart	0.674	0.530	0.857

Figure 21. Odds ratio of comorbidities in determining mortality

The complication Heart disease and accident with COPD has the highest odds for mortality.

3. EFFECT OF COMORBIDITIES ON READMISSION:

Chi-square test for the relationship between readmission of COPD patients and comorbidities

Frequency Col Pct	Table of readmit_flag by complication							
	complication							
readmit_flag	Accident	Heart	Heart Accident	Heart Smoking	Heart Smoking Accident	Smoking	Smoking Accident	Total
0	7438 90.77	21917 91.09	3475 90.40	7621 90.85	1492 91.09	21250 92.01	3578 91.84	66771
1	756 9.23	2145 8.91	369 9.60	768 9.15	146 8.91	1845 7.99	318 8.16	6347
Total	8194	24062	3844	8389	1638	23095	3896	73118

Statistics for Table of readmit_flag by complication

Statistic	DF	Value	Prob
Chi-Square	6	26.5902	0.0002
Likelihood Ratio Chi-Square	6	26.7308	0.0002
Mantel-Haenszel Chi-Square	1	18.2902	<.0001
Phi Coefficient		0.0191	
Contingency Coefficient		0.0191	
Cramer's V		0.0191	

Figure 22. Chi-square test on readmit within 30 days and comorbidities

The Chi-square value is significant. There is association between readmission and Comorbidity.

Tukey's Post-hoc Tests:

Tukey Style Multiple Comparisons of Proportions

Compare	Diff	SE	q	q(.05)	Conclude
6 vs 3	2.08	0.35	5.95	4.17	R
6 vs 1	1.77	0.26	6.81	4.17	R
6 vs 4	1.31	0.26	5.05	4.17	R
6 vs 2	1.21	0.18	6.7	4.17	R
6 vs 5	1.09	0.51	2.14	4.17	A
6 vs 7	0.22	0.35	0.63	4.17	A
7 vs 3	1.86	0.45	4.14	4.17	A
7 vs 1	1.55	0.39	3.97	4.17	A
7 vs 4	1.09	0.39	2.8	4.17	A
7 vs 2	0.98	0.35	2.81	4.17	A
7 vs 5	0.87	0.59	1.48	4.17	A
5 vs 3	0.99	0.59	1.67	4.17	A
5 vs 1	0.68	0.54	1.25	4.17	A
5 vs 4	0.22	0.54	0.4	4.17	A
5 vs 2	0.11	0.51	0.22	4.17	A
2 vs 3	0.88	0.35	2.5	4.17	A
2 vs 1	0.56	0.26	2.17	4.17	A
2 vs 4	0.11	0.25	0.43	4.17	A
4 vs 3	0.77	0.39	1.97	4.17	A
4 vs 1	0.46	0.31	1.48	4.17	A
1 vs 3	0.31	0.39	0.8	4.17	A

Figure 23. Output for Tukey's post hoc test

The complication smoking has an association with Accident, Heart-Accident, Heart-smoking and Heart complications.

ODDS RATIO:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
complication Accident vs Heart	1.068	0.983	1.160
complication Heart Accident vs Heart	1.107	0.991	1.236
complication Heart Smoking vs Heart	1.012	0.931	1.101
complication Heart Smoking Accident vs Heart	0.984	0.831	1.166
complication Smoking vs Heart	0.863	0.810	0.919
complication Smoking Accident vs Heart	0.886	0.786	0.998

Figure 24. Odds ratio of comorbidities in determining readmission

Based on the odds ratio, the complication Heart Accident has the highest odds for readmission and the smoking has the least odds for readmission.

III. SENTIMENT ANALYSIS ON COPD DRUG REVIEWS

The ubiquitous presence of internet has given rise to explosion of data coming from the people. To analyze the functioning of a product or a medicine, we can leverage the experiences of the functioning of the medicine straight out of the Patients words. Understanding the Patients sentiments associated with medications help Doctors to take better care of their patients.

For the purpose of this study, Concept links, Text Clustering, Rule Based models are implemented to understand the most commonly used terms by the patients with COPD

DATA PREPARATION:

Reviews of COPD patients using the drugs in the study are extracted to understand the patient's opinions on the drugs. 400 reviews were collected from patient's using Advair, 204 Patients reviews of Symbicort and 159 user reviews for Spiriva is collected. The reviews are extracted using a web crawler from the below websites:

- Iodine.com
- Drugs.com
- Rxlist.com
- Viewpoints.com
- Druglib.com

CONCEPT LINKS:

Concept linking is a way to find and display the terms that are highly associated with the selected term in the Terms table. The selected term is surrounded by the terms that correlate the strongest with it. The below are the various Concept links generated for various terms.

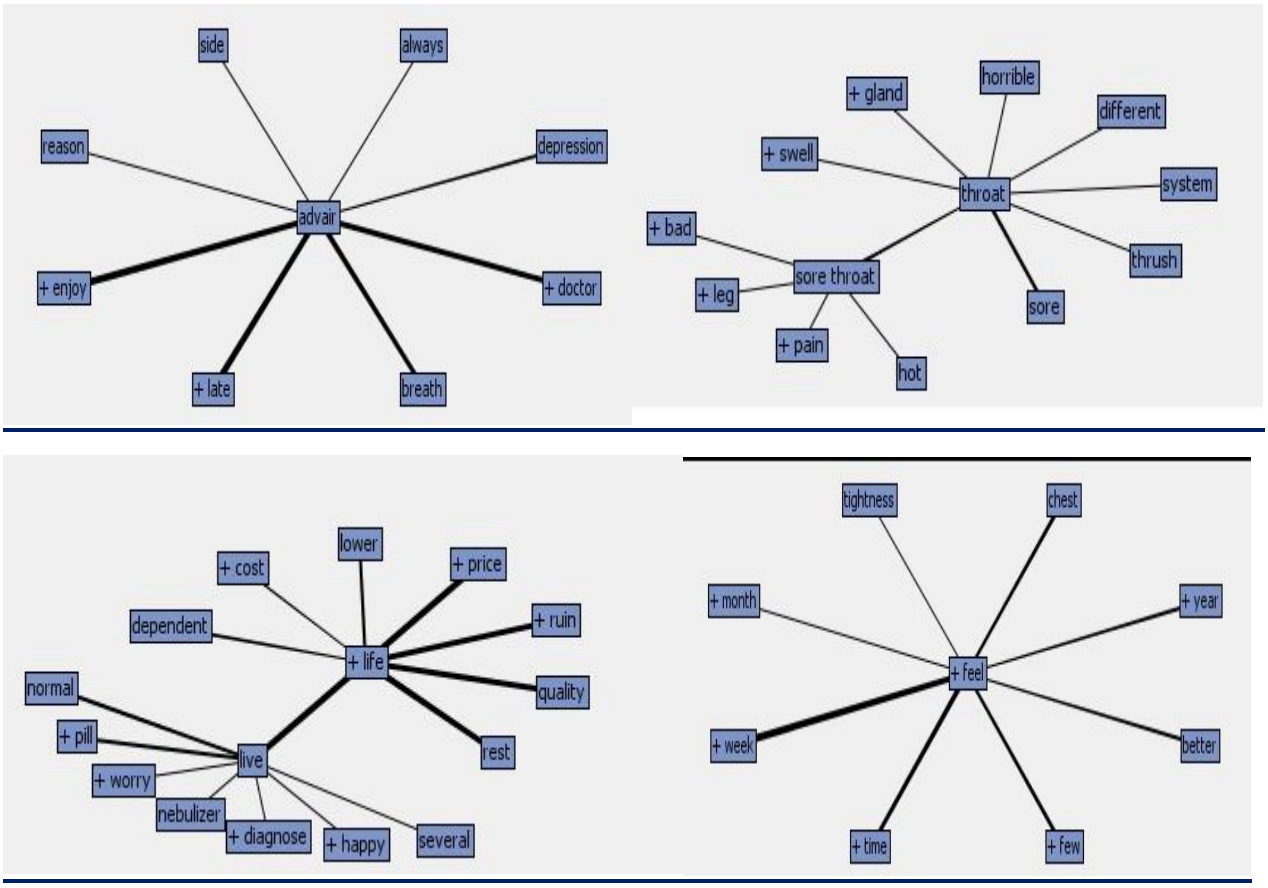


Figure 25. Concept links for Advair drug reviews

The most common terms associated with Advair can be understood through the above concept links. The word Advair is strongly associated with enjoy, breath and doctor. From this concept link, it is understandable that Advair eases the breathing of patients and makes their life easy. In other concept links, Advair is associated with causing symptoms, higher price, increased quality of life etc., The concept links are one of the ways to visualize the association between opinions of the patients.

TEXT CLUSTERING:

The process of dividing a data set into mutually exclusive groups so that the observations for each group are as close as possible to one another and different groups are as far as possible from one another. In SAS Text Miner, clustering involves discovering groups of documents that are more similar to each other than they are to the rest of the documents in the collection. When the clusters are determined, examining the words that occur in the cluster reveals the focus of the cluster. Forming clusters within the document collection can help you to understand and summarize the collection without reading every document. The clusters can reveal the central themes and key concepts that are emphasized by the collection.

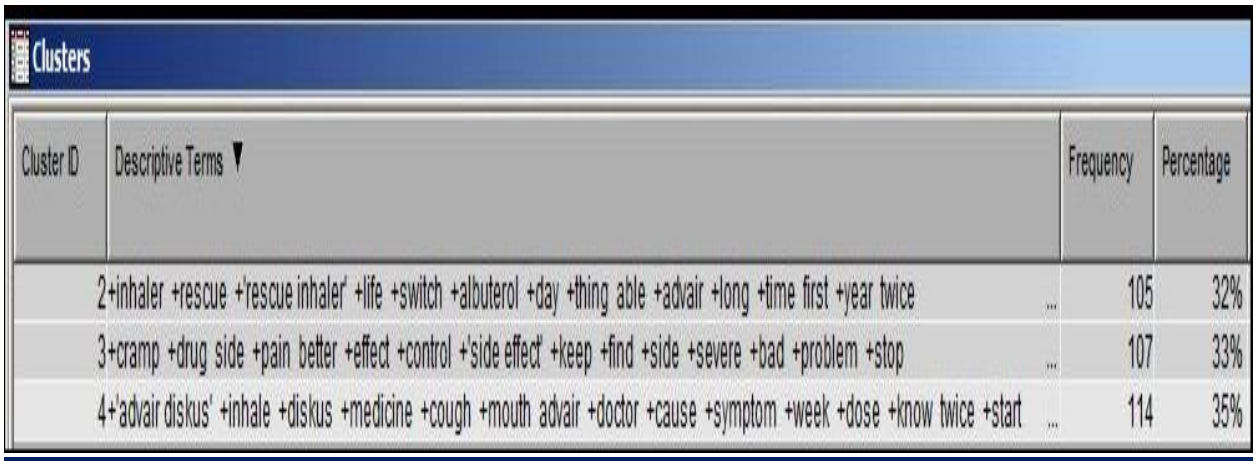


Figure 26. Distribution of Text Clusters

Building Models to Classify Text as Positive/Negative:

SAS Sentiment Analysis Studio provides a way to quickly understand customers'/persons opinions and experiences across multiple channels. The objective of this analysis is to develop a Statistical/rule based model that automatically classifies a given file as Positive/Negative. The below portal is used to classify documents into Positive/Negative/Neutral and allow the Studio to learn from the given data.

After classifying the data, Statistical models can be built to predict the tone of the reviews. In the below model, one of the technique provided an accuracy of 54.55%.

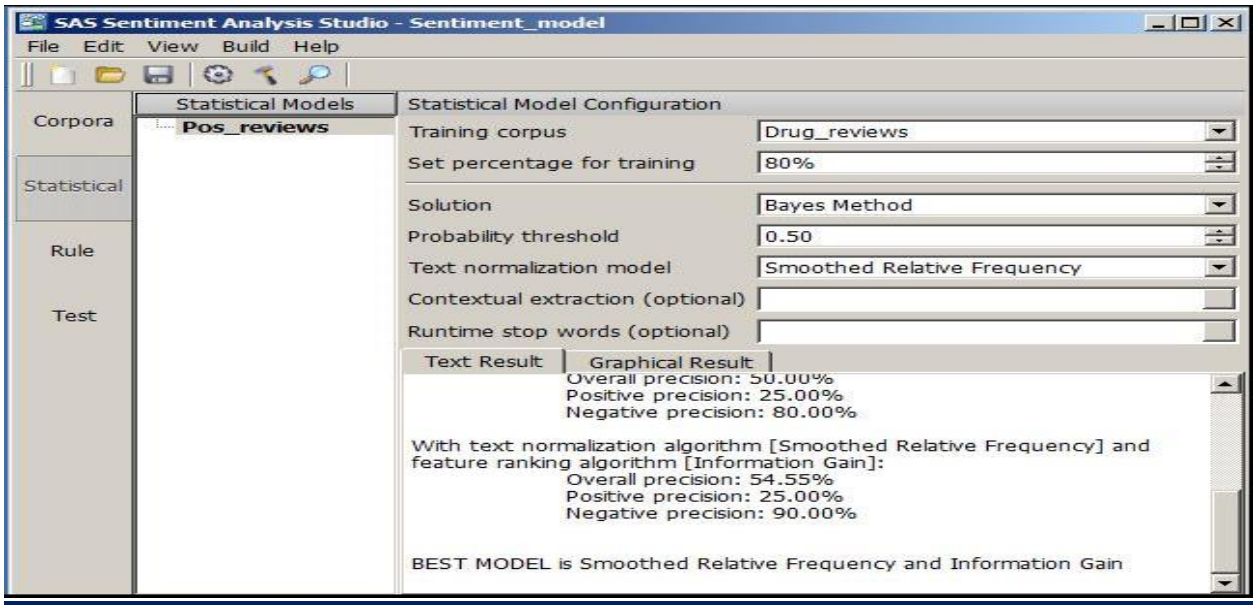


Figure 27. Statistical model output for Advair reviews

RULE BASED MODELS:

This model extracts tokens from the given data files and present the tokens as Positive/Negative or Neutral tone. This gives a quick glimpse of the kind of words used by People when describing about the product.

Positive keywords:

Positive | Negative | Neutral

	Type	Body	Weight
1	CLASSIFIER	run	0.022579
2	CLASSIFIER	wheeze	0.022579
3	CLASSIFIER	medicine	0.017921
4	CLASSIFIER	hope	0.016247
5	CLASSIFIER	notice	0.011378
6	CLASSIFIER	inhaler	0.010725
7	CLASSIFIER	blessing	0.010464
8	CLASSIFIER	level	0.010464
9	CLASSIFIER	bit	0.010464
10	CLASSIFIER	inhale	0.010464
11	CLASSIFIER	love	0.010464
12	CLASSIFIER	Diskus	0.010464
13	CLASSIFIER	life	0.008715
14	CLASSIFIER	mile	0.0069
15	CLASSIFIER	always	0.0069
16	CLASSIFIER	breath	0.0069
17	CLASSIFIER	remember	0.0069
18	CLASSIFIER	period	0.005432
19	CLASSIFIER	wonderfully	0.005432

Figure 28. Commonly used positive words on Advair

Negative keywords:

Positive | Negative | Neutral

	Type	Body	Weight
1	CLASSIFIER	drug	0.072403
2	CLASSIFIER	pain	0.036985
3	CLASSIFIER	problem	0.031182
4	CLASSIFIER	severe	0.029518
5	CLASSIFIER	worse	0.02847
6	CLASSIFIER	leg	0.02847
7	CLASSIFIER	tooth	0.02847
8	CLASSIFIER	cramp	0.022548
9	CLASSIFIER	voice	0.022548
10	CLASSIFIER	long	0.022548
11	CLASSIFIER	night	0.022548
12	CLASSIFIER	advair	0.022548
13	CLASSIFIER	effect	0.021356
14	CLASSIFIER	throat	0.02094
15	CLASSIFIER	side effects	0.02094
16	CLASSIFIER	Use	0.020901
17	CLASSIFIER	Now	0.020901
18	CLASSIFIER	pound	0.020901
19	CLASSIFIER	gain	0.020901

Search Result | Syntax Errors | Rule Evaluation Result

Figure 29. Commonly used negative words on Advair

Model Test:

The model being built in the earlier phase can be used to understand the sentiments of users. This model is used to test the sentiments of reviews of the drugs extracted from the drug portals and categorize into positive/negative.

In the below example, out of 159 user reviews on Spiriva, 54 are categorized as Positive and 105 are categorized as negative articles.

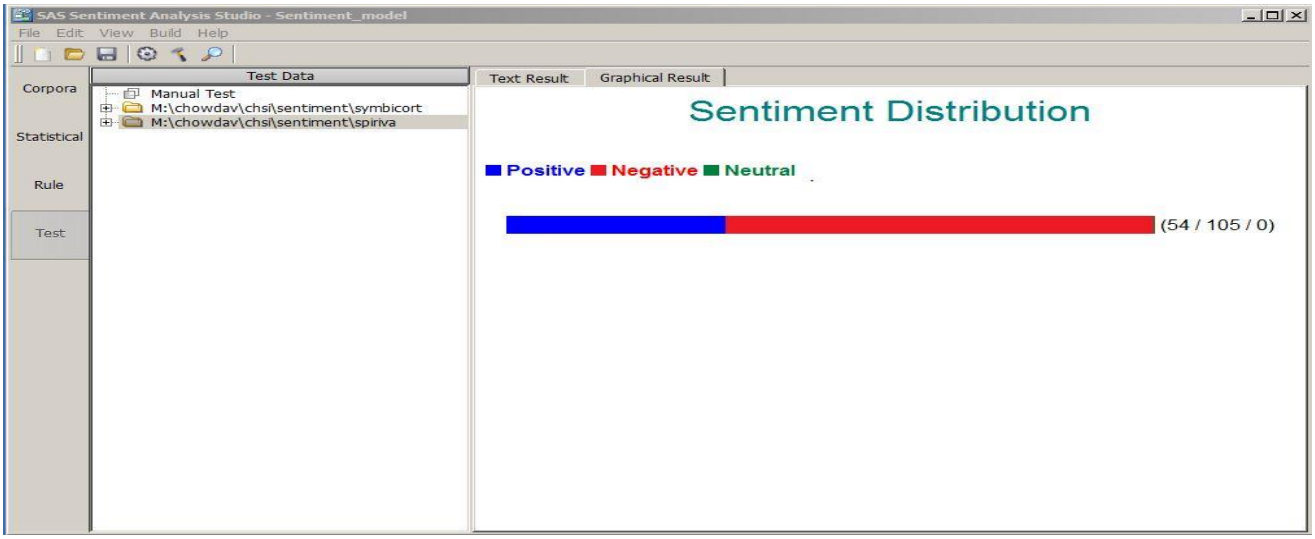


Figure 30. Categorizing the user reviews based on their sentiment

Among Advair, Spiriva and Symbicort, Advair received highest positive reviews with 36.45%, 33.45% viewed Spiriva as positive and only 28% of the user reviews rated Symbicort as positive.

CONCLUSION

EFFECTIVENESS OF COPD DRUGS:

i. Length of Stay:

- The average length of stay for patients with COPD is the least for patients who are prescribed Advair with 5.72 days for over 43197 patient encounters.
- No significant difference between length of stay is observed between Symbicort and Spiriva.
- The combination of Symbicort/Spiriva has the highest avg. length of stay with 6.54 days for over 10706 patient encounters

ii. Readmission within 30 days of discharge:

- The patients who are administered Spiriva has less chances of readmission within 30 days of discharge compared to other drugs.
- Patients using Spiriva has 25% less chances of readmitting compared to Advair, while Symbicort has 8% less chances of readmissions.
- The combination of Symbicort/Spiriva has worst readmission outcome compared to other drugs with 37% more chance of readmission compared to Advair.

iii. Mortality from COPD

- Patients administered with Advair has the best outcomes in terms of mortality from COPD. They are more likely to survive when administered Advair than other drugs.
- Compared to Advair, Spiriva has 19% more chances of mortality, while it is 10% for Symbicort
- The combination of Symbicort/Spiriva has the worst mortality outcome with 31% more chances of mortality compared to Advair.

EFFECT OF COMORBIDITIES ON COPD PATIENTS:

i. Length of Stay:

- The avg. length of stay for patients with COPD is the least for patients who have Smoking disorder with 4.8 days for over 23000 patient encounters.
- The Comorbidity of Accidents with COPD has the avg. length of stay with 7.9 days for over 8500 patient encounters.
- The Patients with Comorbidities of Heart disease and Accidents with COPD has the highest avg. length of stay with 8.9 days for over 4000 patient encounters.

ii. Readmission within 30 days of discharge:

- The Patients with Comorbidities of Heart disease and Accidents has more chances of readmission within 30 days of discharge compared to other drugs.
- Compared to Heart disease, Patients with smoking disorder are 13.7% more likely to survive.
- Overall, there is not a significant influence between different Comorbidities in Patients readmission chances.

iii. Mortality from COPD

- The Patients with Comorbidities of Heart disease and Accidents has the worst outcome in terms of mortality from COPD. They are 85% more likely not to survive compared to single comorbidity of Heart disease.
- Compared to Heart disease, Patients with smoking disorder are 55% more likely to survive.

- Of all the comorbidities, Patients with Smoking disorder are more likely to survive and Accident has more chances of mortality.

SENTIMENT ANALYSIS ON DRUGS:

- Of the three drugs, Advair has received more positive reviews compared to other drugs, while Spiriva has more negative reviews.
- Advair has more positive reviews in terms Asthma maintenance, while Symbicort has better reviews in terms of COPD maintenance.

LIMITATIONS OF THE STUDY

The current study did not include factors such as Patient's condition at the time of admission, Patient's age, COPD stage(severe/mild) etc., The study only examines the effectiveness of the three drugs and their combinations in determining the patient's health outcome, which sometimes may not be sufficient in determining Patient's health outcome.

REFERENCES

1) Kern DM, Davis J, Williams SA, Tunceli O, Wu B, Hollis S, Strange C, Trudo F 2015. "Comparative effectiveness of budesonide/formoterol combination and fluticasone/salmeterol combination among chronic obstructive pulmonary disease patients new to controller treatment: a US administrative claims database study".

2) Lucie Blais, Amelie Forget, Sulabha Ramachandran. 2010. "Relative effectiveness of budesonide/formoterol and fluticasone propionate/salmeterol in a 1-year, population-based, matched cohort study of patients with chronic obstructive pulmonary disease (COPD): Effect on COPD-related exacerbations, emergency department visits and hospitalizations, medication utilization, and treatment adherence".

3) Rudolf Jakob Freund, William J. Wilson, Donna L. Mohr. 2010. *Statistical Methods.* 3rd ed. Elsevier.

4) Alan C. Elliott, Joan S. Reisch. "Implementing a Multiple Comparison Test for Proportions in a 2xc Cross tabulation in SAS®". http://www2.sas.com/proceedings/sugi31/204-31.pdf

5) Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® by Goutam Chakraborty, Murali Pagolu, Satish Garla.

6) Reviews of Advair, Spiriva and Symbicort on drug portals. www.drugs.com, www.rx-list.com, www.druglib.com

ACKNOWLEDGMENTS

We thank SAS Global Forum 2017 conference committee for giving us an opportunity to present our work. We also thank Dr. William Paiva and Ms. Elvena Fong of Center for Health Systems Innovation at Oklahoma State University for granting us permission to publish our results. Special thanks to Dr. Goutam Chakraborty and Dr. Miriam McGaugh of Business Analytics department at Oklahoma State University for their continuous support and guidance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Indra kiran Chowdavarapu
MS in Business Analytics
Oklahoma State University
Phone: 330-461-7511
Email: manikiran.indra@gmail.com
LinkedIN: www.linkedin.com/in/indrakiran

Indra kiran Chowdavarapu is a graduate student enrolled in Business Analytics at the Spears School of Business, Oklahoma State University. He has been working on healthcare analytics projects with the Center for Health Systems and Innovation, Tulsa as a data analytics research assistant since Jan 2016. He is a recipient of SAS Student Ambassador, 2017, highest student honor by SAS Institute. He is a SAS® Certified Advanced Programmer, and also a SAS® Certified Predictive Modeler. He has co-authored a paper presented at the SCSUG conference, San Antonio in 2016 and had a poster presentation at the SAS® Analytics conference, Las Vegas in 2016.