Table of Contents

Executive Summary

Problem Statement

The hypothetical Institute for Improved Community Health (IICH) wishes to find which weather storms, if any, have an impact on the use of health facilities, and if so, IICH would like to have a model to forecast variation of healthcare usage under various assumptions of weather conditions. IICH posits that the frequency of inpatient discharges with Diagnostic Related Group (DRG) codes assigned to infectious diseases might be significantly affected by some kinds of storms.

Specific Activities Required

1. Identify the type of storms that have statistically significant impact on the provided DRG admits.

2. Build a predictive model to predict the number of hospital admissions by DRG, age-group, area code, and week.

Key Findings

We can confirm that certain weather storms have a statistically significant impact on the provided DRG admit frequency. The table below provides a quick summary of the statistically significant combinations (cells marked in green), which were derived using Market Basket Analysis.

	DRG Codes Grouped by Associated Age Levels																						
	0 - 17					18+										All ages							
	70	81	91	98	422	68	69	79	80	89	90	96	97	419	420	421	21	76	99	100	101	102	423
Storm Type																							
WS			■	■				■		■		■	■					■					
TS										■													
W										■													
C			■	■				■		■		■	■					■					
F			■	■				■		■													

The basic concept of the association of infectious disease to weather is the belief that such a disease is much more likely to be spread in the community when people spend more time indoors in close contact with each other.

Winter Storms (WS) and Cold Storms (C) have the most number of associations, impacting 7 of the 23 DRG admit types. These storms have the most severe impact on the weather conditions, causing large drops in temperature, high levels of snowfall and a reduction in the average amount of daylight. Our findings therefore match the basic concept of association- Winter and Cold Storms result in cold temperatures and snow, which increases the amount of time spent indoors leading to a higher transmission of infection.

Flood, and to a lesser degree Thunderstorm and Windstorm are also linked with certain DRG admits. These weather events are associated with very high levels of precipitation, which tend to increase the time spent indoors, and hence increase the chances of disease transmission.

Our Predictive model for hospital admissions is constructed as a separate model for Ages 0-17, and a separate model for Ages 18-65+. Both models use a Feed Forward Neural Network modelling algorithm, with input selection provided by a stepwise logistic regression. When assessed using a validation sample the models achieve an average squared error of 0.2065 and 0.2463 for the 0-17 and 18-65+ year olds respectively.

To arrive at our predictive model for admit frequency, we tested a wide variety of statistical models and algorithms including Decision Trees, Neural Networks, Logistic Regression, and a variety of Generalized Linear Models (Poisson, Negative Binomial, Zero Inflated Poisson). We also tried a variety of target variables, and a selection of different input variables. Candidate models were assessed on their ability to predict the number of admits in a holdout sample- a sample of data not used in the training phase.

The Decision Tree model trained on the admit counts produced very strong candidate models, and are used to validate the Market Basket Analysis results. In order of importance the variables most associated with hospital admits are

Ages 0-17:	Population, Age, DRG Group, Seasonality, **Winter Storm**, **Cold Storm**, Schooldays
Ages 18-65:	Population, Age, DRG Group, **Winter Storm**, Seasonality, **Cold Storm**, Workdays, **Wind Storm**

Flood and Thunderstorm were not considered important to predicting admit frequency- the other available inputs did a better job.

Review of the literature

Knowledge of the interactions between climate and health dates back to the time of Aristotle (384–322 BC), but our understanding of this subject has recently progressed rapidly as technology has become more advanced. At the same time, our ability to forecast weather and climate (in terms of both accuracy and lead-times) has improved significantly in recent years. The increased accuracy of climate predictions, and improving understanding of interactions between weather and infectious disease, has motivated attempts to develop models to predict changes in the incidence of epidemic-prone infectious diseases. (*World Health Organization 2005: Using Climate to Predict Infections disease epidemics*)

The direct impact of climate on infectious diseases can occur by three principal pathways: 1) effects on human behavior; 2) effects on the disease pathogen; and 3) effects on the disease vector.

Human Behavior

One key component of the association of infectious disease to weather is the belief that such a disease is much more likely to be spread in the community when people spend more time indoors in close contact with each other. The strong seasonal pattern of influenza infections in Europe, for example, is thought to reflect the increased tendency among humans to spend more time indoors during the winter months (*Halstead, 1996*). It is also believed that school and work provide environments with much social contact that is somewhat mitigated by good weather when people 'get out' much more.

Disease pathogens

For infectious diseases caused by a pathogen that develops outside the human host (i.e. in the environment or in an intermediate host or vector), climate factors can have a direct impact on the development of the pathogen. Most viruses, bacteria and parasites do not complete their development if the temperature is below a certain threshold (*e.g. 18 ºC for the malaria parasite Plasmodium falciparum and 20 ºC for the Japanese encephalitis virus; Macdonald, 1957; Mellor & Leake, 2000*). Increases in ambient temperature above this threshold will shorten the time needed for the development of the pathogen and increase reproduction rates, whereas temperatures in excess of the tolerance range of the pathogen may increase mortality rates.

Disease vectors

The geographical distribution and population dynamics of insect vectors are closely related to patterns of temperature, rainfall and humidity. A rise in temperature accelerates the metabolic rate of insects, increases egg production and increases the frequency of blood feeds (*e.g. Detinova, 1962; Mellor & Leake, 2000*). The influence of rainfall is also often significant, although it is less easy to predict. Rainfall has an indirect effect on vector longevity through its effect on humidity; relatively wet conditions create favorable insect habitats and thereby increase the geographical distribution and seasonal abundance of disease vectors. In other cases excess rainfall may have catastrophic effects on local vector populations if flooding washes away the breeding sites.

Problem Understanding

The timing of the weather events and hospital admits is a key factor in establishing a link between the two:

"Because the development of a disease requires an incubation period, each person with an 'infectious' exposure might be expected to need the healthcare service at least some days after a storm; however, after allowing some additional time for the symptoms to worsen, any time beyond that would force disassociation of the weather event from any resultant admit. Thus, for each DRG, we assign a minimum incubation period and a maximum time for association of the weather event and the admission."

A storm event can only be associated with resulting DRG admits, that occur within a fixed time period after the storm. Consider DRG code 420, which has a minimum incubation period of 1 week, and a maximum time for association of 2 weeks.

DRG code	week	Storm Event	Association Period	Admit
420	262	0	0	0
420	263	0	0	0
420	264	0	0	0
420	265	1	0	1
420	266	0	1	0
420	267	0	1	1
420	268	0	0	0
420	269	0	0	0
420	270	0	0	0

The storm occurs in week 265, giving us an association period of weeks 266 and 267. Any admits occurring outside of this period cannot be associated with that storm. Every DRG has a minimum incubation period of 1 week. The maximum time for association varies between 2 and 4 weeks, depending on the DRG.

It is also realized that the socialization habits and immunity systems of people tend to change depending of phases of biological development. Patient ages and population data in eight segments is available and should be incorporated into the solution.

Our objective is to find rules with quantifiable confidence and lift to determine how important certain kinds of storms might be responsible for the identified DRG admit frequency jumps. Different DRGs might be more related to some types of storms than others – and some storms might not be statistically associated with any DRGs.

To complete the required activities specified in the problem statement, we complete two analysis tasks.

Analysis Task 1: Identify the type of storms that have statistically significant impact on the provided DRG admits.

Analysis Task 2: Build a predictive model to predict the number of infectious diseases by DRG, age-group, area code, and week in the Score Dataset

The results of Analysis Task 1 will inform approaches taken in Task 2, and some results from Task 2 will be compared back to those of Task 1 for the purposes of validation.

Data Understanding

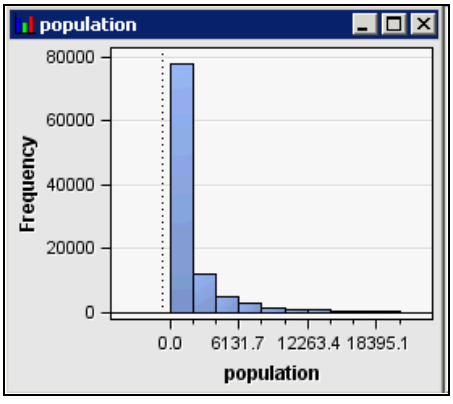
Quick View

721,808 records were available for initial investigation. 190,772 (26%) records had an Admit count ≥ 1, while 531,036 (74%) had Admits = 0.Below is a table that shows the total number of admits by eight Age Groups.

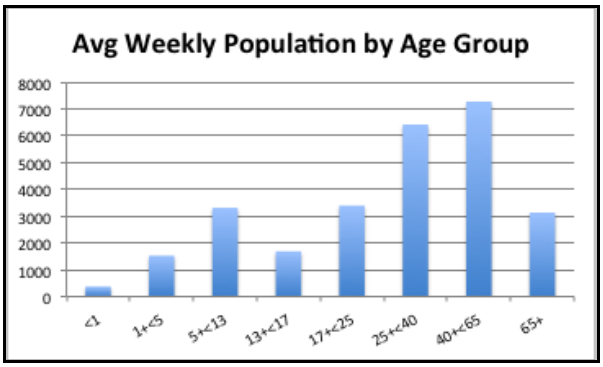
	Number of Admits	Percent
Age Group		
<1	25093	12%
1+<5	12822	6%
5+<13	9479	5%
13+<17	3682	2%
17+<25	6469	3%
25+<40	15182	7%
40+<65	51088	24%
65+	84808	40%
Total	208623	100%

Population

Population is an important component driving the number of admits. All things being equal, a larger sub-population will have higher Admit counts than a smaller sub-population. The records in our data set represent Admit counts for a week, in a given Area Code, for a given age group. The provided population numbers are also at this level. The minimum weekly population for an Age Group and Week is 0, while the maximum is 20,439. The population distribution, with a heavy right skew, is shown below.

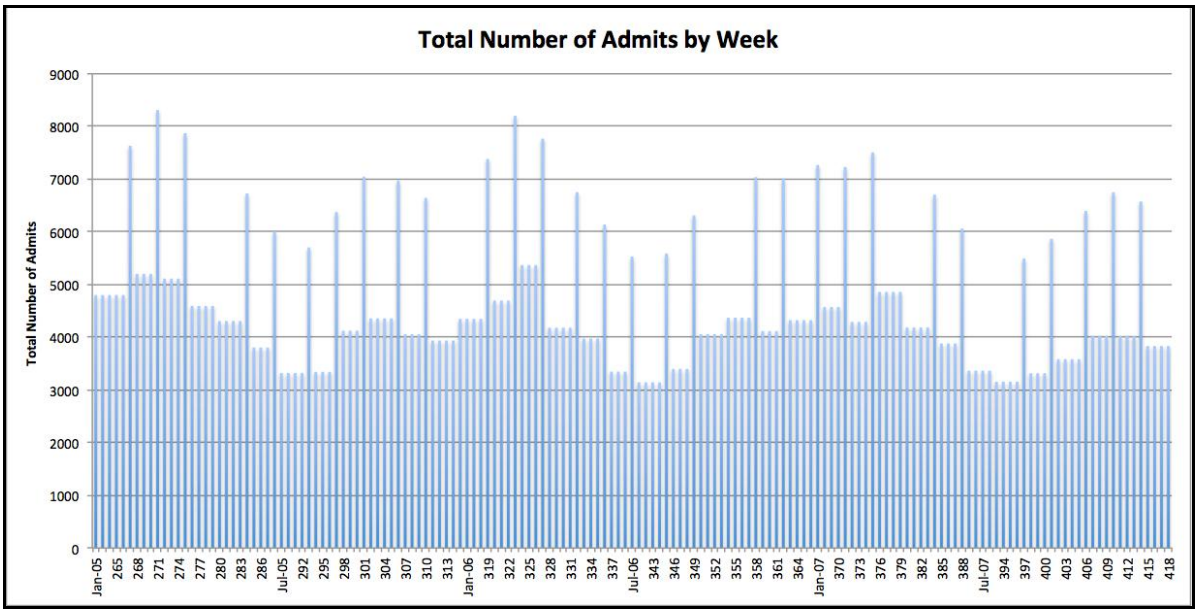


The weekly population for an Age group can vary widely between area codes. For example, the age group "1+ < 5" has an average weekly population of 1,325 in area code VUX6VZ, while in area code OOL71RS it is only 242. Similar variation is found at all Age Levels. As mentioned, the age groups have differently sized populations. The chart below shows the average weekly population across all area codes by age group.



Admits

The literature review revealed that many infectious diseases follow a seasonal or cyclical pattern, Our DRG admit data also exhibit this pattern, with peaks occurring in the winter months.

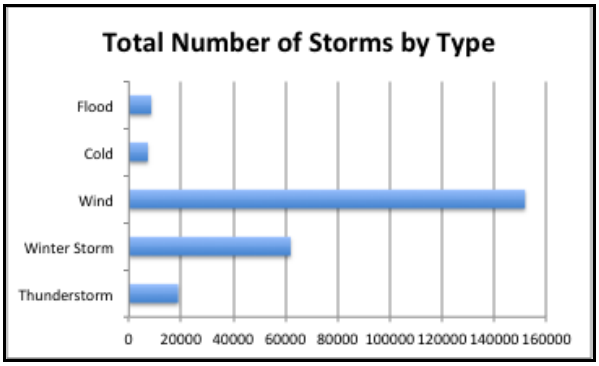


The table below shows the average probabilities of a record having a non-zero admit, for different age groups. The very young and very old are most at risk of infection.

	Admit = 0	Admit ≥ 1	N	P (Admit ≥ 1)
Age Group				
<1	61680	22948	84628	0.27
1+<5	40427	12553	52980	0.24
5+<13	30782	9268	40050	0.23
13+<17	13780	3676	17456	0.21
17+<25	24436	6461	30897	0.21
25+<40	54873	15071	69944	0.22
40+<65	142412	48107	190519	0.25
65+	162646	72688	235334	0.31
Total	531036	190772	721808	0.26

Storms

During the 3-year period in our data represents, the weekly number of five storm types was tracked. Wind is the most frequently occurring storm, followed by Winter Storm and Thunderstorm.



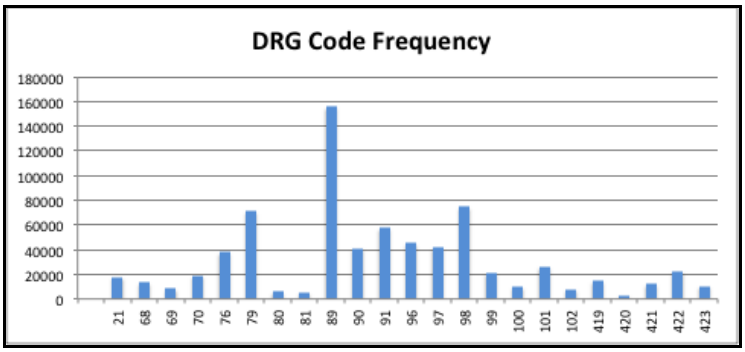
DRGs

DRG codes track the type of infectious disease a patient presents with. The table below shows the 23 different DRG codes in the data, and how they map to different age levels. 5 apply to patients that are 17 years of age and under. 12 are for patients over 17 years of age. And 6 codes apply to all age levels.

DRG Codes by Age Range Classification

17 and Under	Over 17
70: OTITIS MEDIA & URI AGE 0-17	68: OTITIS MEDIA & URI AGE >17 W CC
81: RESPIRATORY INFECTIONS & INFLAMMATIONS AGE 0-17	69: OTITIS MEDIA & URI AGE >17 W/O CC
91: SIMPLE PNEUMONIA & PLEURISY AGE 0-17	79: RESPIRATORY INFECTIONS & INFLAMMATIONS AGE >17 W CC
98: BRONCHITIS & ASTHMA AGE 0-17	80: RESPIRATORY INFECTIONS & INFLAMMATIONS AGE >17 W/O CC
422: VIRAL ILLNESS & FEVER OF UNKNOWN ORIGIN AGE 0-17	89: SIMPLE PNEUMONIA & PLEURISY AGE >17 W CC
	90: SIMPLE PNEUMONIA & PLEURISY AGE >17 W/O CC
All Ages	96: BRONCHITIS & ASTHMA AGE >17 W CC
	97: BRONCHITIS & ASTHMA AGE >17 W/O CC
21: VIRAL MENINGITIS	100: RESPIRATORY SIGNS & SYMPTOMS W/O CC
76: OTHER RESP SYSTEM O.R. PROCEDURES W CC	419: FEVER OF UNKNOWN ORIGIN AGE >17 W CC
99: RESPIRATORY SIGNS & SYMPTOMS W CC	420: FEVER OF UNKNOWN ORIGIN AGE >17 W/O CC
101: OTHER RESPIRATORY SYSTEM DIAGNOSES W CC	421: VIRAL ILLNESS AGE >17
102: OTHER RESPIRATORY SYSTEM DIAGNOSES W/O CC	
423: OTHER INFECTIOUS & PARASITIC DISEASES DIAGNOSES	

The chart below shows the relative frequency of DRG codes. DRG 89, simple pneumonia is by far the most common diagnosis in the data. Codes 76, 79, 90, 91, 96, 97, and 98 have moderate frequency, while the others are quite low. DRG 420 is the most uncommon.



Missing Data

In combining the data tables we observed varying degrees of incomplete or missing data. For example the admit table contains data for the 23 DRG codes, at an area code and week level. There appear to be gaps in the reporting period however. Consider the following table, which shows admit data for age group (65+), area code (ZZE19WE), and DRG code (89).

	Area Code	DRG24	age_group	week	admits	week_area_age	age_group2
4796	ZZE19WE	89	65+	262	0	262ZZE19WE65+	h. 65+
9590	ZZE19WE	89	65+	263	1	263ZZE19WE65+	h. 65+
14386	ZZE19WE	89	65+	264	0	264ZZE19WE65+	h. 65+
19183	ZZE19WE	89	65+	265	0	265ZZE19WE65+	h. 65+
23978	ZZE19WE	89	65+	266	0	266ZZE19WE65+	h. 65+
31601	ZZE19WE	89	65+	267	1	267ZZE19WE65+	h. 65+
36799	ZZE19WE	89	65+	268	0	268ZZE19WE65+	h. 65+
41995	ZZE19WE	89	65+	269	0	269ZZE19WE65+	h. 65+
47191	ZZE19WE	89	65+	270	1	270ZZE19WE65+	h. 65+
55491	ZZE19WE	89	65+	271	0	271ZZE19WE65+	h. 65+
78676	ZZE19WE	89	65+	275	0	275ZZE19WE65+	h. 65+
83266	ZZE19WE	89	65+	276	1	276ZZE19WE65+	h. 65+
87854	ZZE19WE	89	65+	277	0	277ZZE19WE65+	h. 65+
92441	ZZE19WE	89	65+	278	0	278ZZE19WE65+	h. 65+
97029	ZZE19WE	89	65+	279	0	279ZZE19WE65+	h. 65+
138368	ZZE19WE	89	65+	288	0	288ZZE19WE65+	h. 65+
141690	ZZE19WE	89	65+	289	0	289ZZE19WE65+	h. 65+
145009	ZZE19WE	89	65+	290	1	290ZZE19WE65+	h. 65+
148328	ZZE19WE	89	65+	291	0	291ZZE19WE65+	h. 65+
151647	ZZE19WE	89	65+	292	0	292ZZE19WE65+	h. 65+
157344	ZZE19WE	89	65+	293	0	293ZZE19WE65+	h. 65+
160684	ZZE19WE	89	65+	294	0	294ZZE19WE65+	h. 65+
164023	ZZE19WE	89	65+	295	0	295ZZE19WE65+	h. 65+
167362	ZZE19WE	89	65+	296	1	296ZZE19WE65+	h. 65+
173729	ZZE19WE	89	65+	297	0	297ZZE19WE65+	h. 65+
236296	ZZE19WE	89	65+	310	0	310ZZE19WE65+	h. 65+
240230	ZZE19WE	89	65+	311	1	311ZZE19WE65+	h. 65+
244163	ZZE19WE	89	65+	312	0	312ZZE19WE65+	h. 65+
248096	ZZE19WE	89	65+	313	0	313ZZE19WE65+	h. 65+
252029	ZZE19WE	89	65+	314	0	314ZZE19WE65+	h. 65+

Missing data occurs in the other tables to a varying degree. Before we proceed with a quantification it is useful to make the distinction between different the three types of missing data.

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing not at Random (MNAR)

When we say that data are missing completely at random, we mean that the probability that an observation (X_i) is missing is unrelated to the value of X_i or to the value of any other variables.

For data to be missing *completely* at random, the probability that X_i is missing is unrelated to the value of X_i or other variables in the analysis. But the data can be considered as missing at random if the data meet the requirement that missingness does not depend on the value of X_i *after controlling for another variable.* It is unlikely that we would have such variables in our dataset.

If data are not missing at random, or completely at random, then they are classed as MNAR. When we have data that are MNAR we have a problem, which we need to overcome. For example, the gap in the reporting period seen in the admit table may due to inconsistencies in recording zero admits. Hospitals in one state may record zero admits as a zero, others may just not enter a records. Another example could be if population data were systematically missing for the area codes with the highest admit rates, or highest occurrence of storms.

Quantification and Approach

Admits
We assume that gaps in the recording period occur completed at random. Therefore no imputation is required. The admit table is our starting point for joining the tables. For the remaining tables, missing records will be relative to the 722k records in the admit table.

Weather data
Of the 722k admit records, only 421k have weather information. A number of options are available to us

1. Throw away the 301k records with no weather information
2. Impute the missing weather data
3. Only use the weather data in Predictive Modeling algorithms that can handle missing values (Decision Trees)

First however we must ascertain whether our data is missing completely at random. This would be a desirable property, as it would allow us to select option 1 without biasing our results. To test this assertion, we use a Chi-square test of independence. We construct a 2x2 Contingency table to test the following hypotheses:

Ho: Admits are independent of Weather data presence
Ha: Admits are associated with Weather data presence.

	Total Admit Records	721,808
	Weather Missing	Weather Available
1+ admits	81,659	109,113
0 admits	219,590	311,446

For a 2x2 Contingency Table the chi-square statistic is calculated as follows:

	Variable 1		
Variable 2	Data type 1	Data type 2	Totals
Category 1	a	b	a + b
Category 2	c	d	c + d
Total	a + c	b + d	a + b + c + d = N

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)}$$

The statistic for the weather data is 122 which compared to a Chi Square distribution with 1 degree of freedom is highly significant ($p << 0.01$).

We therefore reject the null hypothesis that the weather missingness is independent of admit frequency. Those records with missing weather data have a higher admit rate.

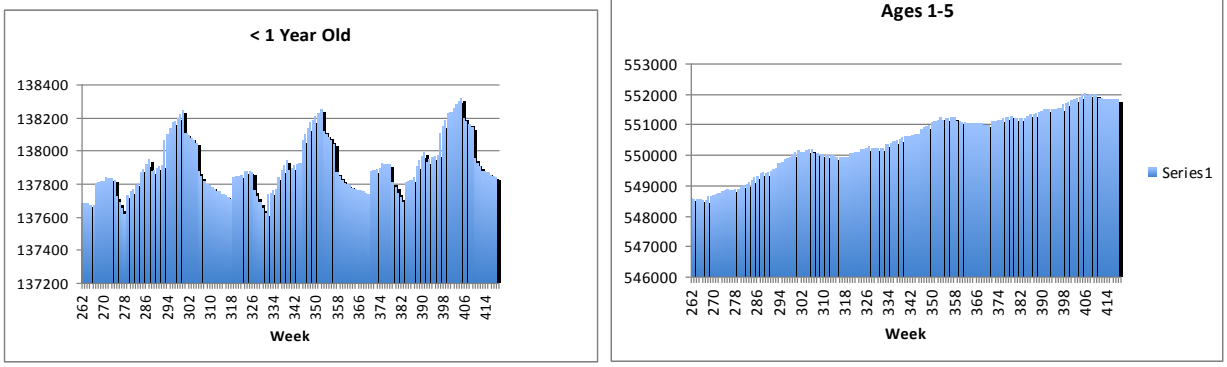
weather	admits	population	rate
AVAILABLE	120247	1723608286	69.7647
MISSING	88376	1184458382	74.6130

This means that option 1 is no longer viable, as it would introduce bias into our results. We also ruled out missing value imputation for two reasons. The weather is extremely variable from day to day, and it is unlikely that our imputations would match the actual weather observed. Ultimately we did not think it sensible to include 7 weather variables that were 42% of the time complete guesses!

We decide that option 3 is the most sensible approach.

Population data

Of the 722k admit records 11k had missing population data. Some of these records did not have a corresponding population entry for any week, leaving us with little clue to the population in these areas. We decided that it would be reasonable to impute these records based on the weekly average populations for each age group. We went to the age group level, as the trend pattern across age group varies- see below.



Outliers

The Admit target variable has an extremely right-skewed distribution. 98% of the records have a 0 or 1, the vast majority of these being 0. Admits over 3 are very rare, but Admits counts of 8 or more also exist in the data.

admits	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	531036	73.57	531036	73.57
1	177073	24.53	708109	98.10
2	10926	1.51	719035	99.62
3	1944	0.27	720979	99.89
4	545	0.08	721524	99.96
5	167	0.02	721691	99.98
6	64	0.01	721755	99.99
7	18	0.00	721773	100.00
8	15	0.00	721788	100.00
9	5	0.00	721793	100.00
10	4	0.00	721797	100.00
11	3	0.00	721800	100.00
12	3	0.00	721803	100.00
13	3	0.00	721806	100.00
14	2	0.00	721808	100.00

Outliers often have ill effects on many predictive models, therefore we expected to remove the outliers so models could focus on estimating the core of the data. However, upon further investigation, removing outliers did not improve model performance. In fact, it worsened performance slightly. Multiple models were used to test different truncations of Admit values. In the end, no records were removed from analysis because of extreme admit counts.

Storm Profiling

Although the weather has not been used as an input variable for the vast majority of our modelling algorithms, it is now used to derive insight into the nature of the different Storm types. It is hoped that this insight, together with the results of the association analysis and predictive modelling, leads us to a wider understanding of the effect of Storms on admit frequency.

Storm data is available for 5 different types of storm.

Group	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cold	1736	4.06	1736	4.06
Flood	2101	4.92	3837	8.98
Thunderstorm	2056	4.81	5893	13.79
Wind	19251	45.05	25144	58.84
Winterstorm	17589	41.16	42733	100.00

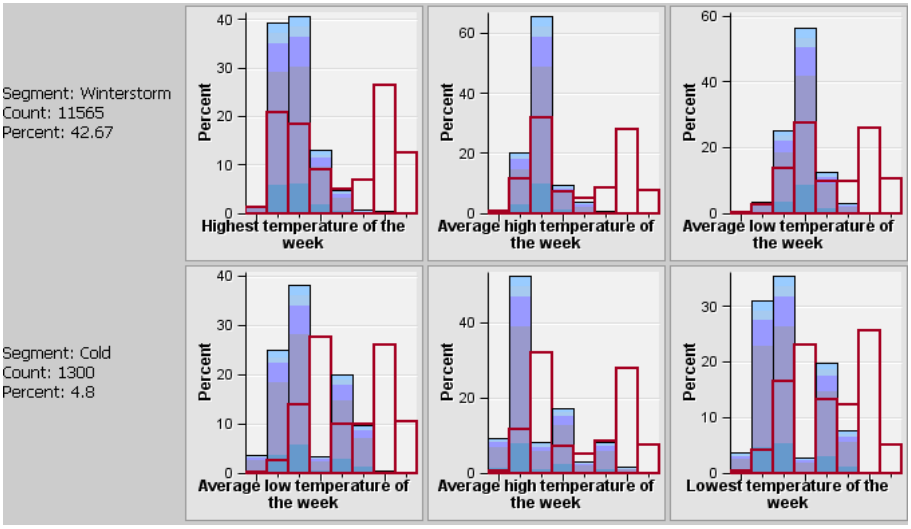
Wind and Winterstorm are the most common, accounting for 86% of all storms. Cold, Flood and Thuderstorm account for the remaining 14%.

So that we can better understand the storms and their effect on the weather, we attach the weather table. Unfortunately 37% of the storm records do not have corresponding weather information.

weather	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AVAILABLE	27101	63.42	27101	63.42
MISSING	15632	36.58	42733	100.00

However we decide to work with the information available, and use the 27k storm records with matching weather data in our analysis. Our objective is to profile the different storm types, using the weather information (average low temperature, rainfall) as descriptors.

Winter storm and Cold Storm produce extreme changes to the weather. All 4 temperature measurements are significantly lower than the base line. Additionally the levels of snow fall are much higher than the baseline. These weather conditions are very likely to drive people to congregate inside, and potentially lead to an increase in disease transmission.



The blue bars relate to the specific weather storm, and the red bars relate to the weather conditions of a typical storm.

Flood and to a lesser degree Wind and Thunderstorm all increase the levels of precipiation dramatically. People are more likely to remain indoors during these weather conditions.

Interestingly the Average LotT and MinLowT for Thunderstorms are around 50% higher than the baseline. This change in weather may effect the disease pathogens/vecotors leading to more disease infections.

Actual Readings

	AvgLowT	MinLowT	AvgHighT	MaxHighT	prcp	snow	Daylight
No Storm	38.8	30.8	57.3	66.3	0.5	0.2	11.8
Winterstorm	18.8	10.4	31.4	37.8	0.7	2.0	9.4
Wind	51.2	42.5	72.3	82.2	1.0	0.1	13.4
Flood	43.9	35.7	63.7	74.9	1.2	0.0	12.6
Thunderstorm	60.0	53.2	80.5	88.2	0.9	0.0	14.0
Cold	18.9	9.9	34.3	40.6	0.7	0.7	9.6

Percentage Differences

	AvgLowT	MinLowT	AvgHighT	MaxHighT	prcp	snow	Daylight
Winterstorm	49%	34%	55%	57%	127%	879%	80%
Wind	132%	138%	126%	124%	194%	26%	114%
Flood	113%	116%	111%	113%	219%	16%	107%
Thunderstorm	155%	173%	141%	133%	169%	0%	119%
Cold	49%	32%	60%	61%	133%	291%	82%

Visual Display

	AvgLowT	MinLowT	AvgHighT	MaxHighT	prcp	snow	Daylight
Winterstorm	low	low	low	low	-	high	-
Wind	-	-	-	-	high	low	-
Flood	-	-	-	-	high	low	-
Thunderstorm	high	high	-	-	high	low	-
Cold	low	low	low	low	-	high	-

Data Preparation

Potential analysis variables were spread out across multiple tables, many of which contained extraneous data (and sometimes not enough data). Our initial data preparation task was to combine all variables into one table that would then be carried forward. The Storm table needed to be transformed into a more condensed, "wider" table so that each type of storm could be accounted for per Admit record. Base SAS was used extensively, and especially SQL Joins for combing tables, records.

After generating a combined table that represented all of the data in the various tables, our next step was to generate user-defined variables thought to be important to subsequent analysis.

One of the most important efforts was in generating Storm Type Counts and binary Indicators for each record. These were created with the specific DRG incubation periods in mind and were the core inputs for our association analysis. For example, if at least one Thunderstorm occurred during the incubation period for the DRG, the Thunderstorm indicator variable would = 1, 0 otherwise

Some variables (Year, Month, and Quarter) were added to reflect the apparent cyclical or seasonal effects apparent in the data. The larger number of DRG levels were also of concern.

First, there is the data dimensionality consideration of having a nominal variable with 23 levels. Second, the DRG codes seemed collapsible based on subject matter knowledge alone.

For example, many DRGs seem to come in pairs based on whether or not the patient presents with complications or not. DRG Codes originated in the 1980s as part of the US Medicare system to control costs (*Gottlober, 2001*). In the cases where a DRG is "doubled" due to clinical complications or lack thereof, this is a direct result of complications driving up costs. This is not a distinction about the nature of the infectious disease that presented in the patient. Therefore, "doubled" DRGs can be collapsed into a single DRG. We created a nominal variable for this purpose, which reduced the 23 levels down to 16 levels.

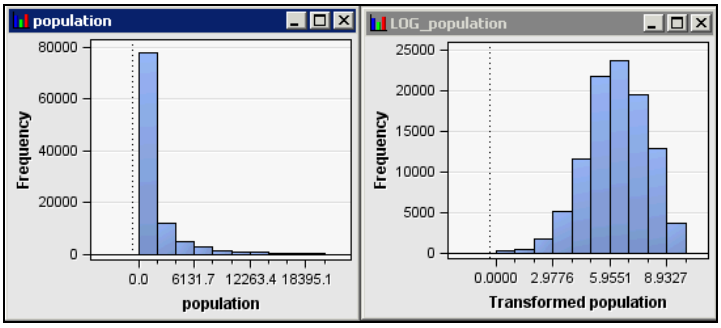
We also used a data driven collapsing method. Via a Decision Tree analysis of DRG against Admits, our collapsed DRG variable resulted in 3 levels of the 17 and under model, and 4 levels for the 18 and over model.

We removed 45 records that had populations of 0, some of which had admits greater than the population.

Two additional Target variables were created for subsequent model investigation:

- Rate - the average weekly admit rate (total admit count / total pop)
- Binary - the average probability of an admit being ≥ 1

When using the binary target, the highly right skewed population variable was transformed before inclusion into models. The resulting Log(pop) transformation created a much improved, more normal distribution as seen below.



An indication of how the two additional targets are related to a few of the key modeling inputs is shown below.

Input Variable	Rate per 10K	Probability	Input Variable	Rate per 10K	Probability
DRG Code			Age Group		
21	0.45	0.21	< 1	8.04	0.27
68	0.37	0.21	1+ < 5	1.61	0.24
69	0.38	0.21	5+ < 13	0.72	0.23
70	1.39	0.21	13+ < 17	1.27	0.21
76	0.40	0.22	17+ < 25	0.63	0.21
79	0.62	0.27	25+ < 40	0.34	0.22
80	0.44	0.21	40+ < 65	0.37	0.25
81	1.38	0.21	65+	1.17	0.31
89	1.17	0.38			
90	0.41	0.21	Winter Storm		
91	1.76	0.23			
96	0.40	0.22	0	0.64	0.25
97	0.36	0.22	1	1.12	0.35
98	2.25	0.29			
99	0.39	0.2	Thunderstorm		
100	0.52	0.21			
101	0.38	0.2	0	0.74	0.27
102	0.42	0.21	1	0.48	0.23
419	0.35	0.2			
420	0.41	0.22	Wind		
421	0.39	0.21			
422	1.76	0.21	0	0.79	0.27
423	0.39	0.21	1	0.53	0.23
Flood			Cold		
0	0.72	0.26	0	0.70	0.26
1	0.70	0.3	1	1.07	0.38

* Storm Types are indicator variables. 1 = present during incubation period, 0 = not present during incubation period

On a one way basis Winter Storm and Cold storm are clearly related to an uplift in admit probability and admit rate.

Finally, our approach to Analysis Task 1 required that parts of the data be transformed from a "wide" organization to a "narrow" or transaction-based view. This was accomplished inside EM via Base SAS and SQL statements.

Methodology & Results

Analysis Task 1: The Effect of Storms on Admission Counts

As previously mentioned, it is known that the disease transmission of certain communicable diseases increases in socially dense situations, such as when many individuals are in close contact at work or school. In addition to this, weather patterns, and specifically storms, often influence group social patterns. For example, it is reasonable to expect people to spend relatively more time indoors during periods of inclement weather.

Our data set includes storm information that is matched to DRG and hospital admission counts. With this information, we can investigate whether or not certain types of diseases (DRGs) show increases in admissions when storms have occurred in the recent past during the incubation period related to the disease.

Association Analysis

One approach to measuring the association (or lack there of) between storms, infectious diseases, and hospital admissions is to use a technique known as Market Basket Analysis (MBA). MBA is a popular analysis tool in the retail sector. For example, by considering the purchases of shoppers -- or the contents of their market baskets -- researchers may be able to discover patterns of particular items occurring together in baskets much more frequently than expected. The results of MBA yield three important metrics: Support, Confidence, and Lift.

As an example, if we are interested in Soda and Popcorn, and in particular, if the presence of Soda increases the likelihood that Popcorn is also present (purchased), we might discover that 10% of all transactions include Soda and Popcorn. Support is 10%. We might learn that of all the transactions that include Soda, Popcorn is present 30% of the time. The Confidence of Soda implying Popcorn is 30%. Finally, if we only expect Popcorn to be present in 20% of all transactions, then Soda is providing a Lift of 1.67 (30% ÷ 20%). The rule "Soda => Popcorn" appears to be an interesting rule. The presence of Soda increases the likelihood that Popcorn is also present.

Given our current analysis task, we can consider DRGs, Storms, and Admits as "items" that might be present in any given transaction. In this case, we are not interested in Age levels or Work Days or any other variables. They are left out of the analysis.

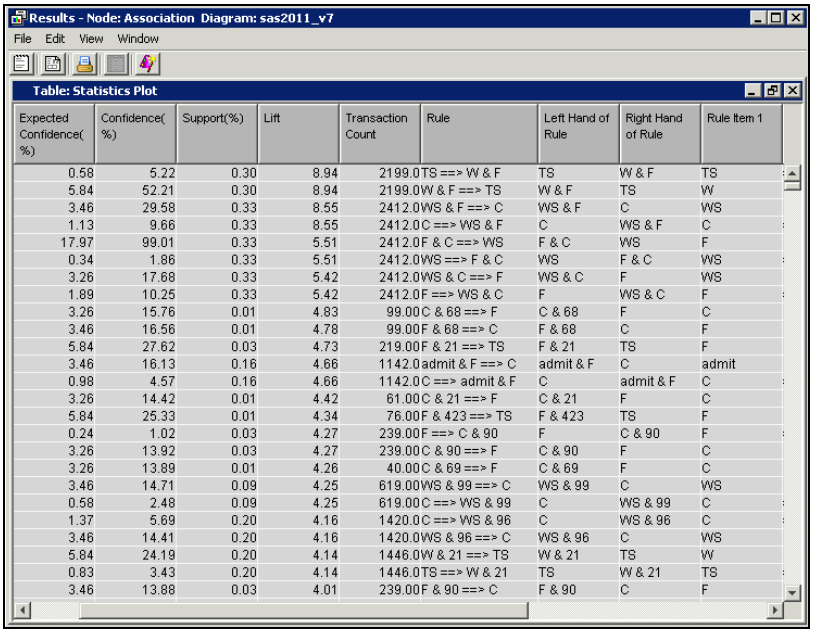
We consider each record in our data set to be a unique transaction. Here are a few possible transactions:

- DRG = 89, Wind storm during incubation period, Hospital Admission
- DRG = 76
- DRG = 79, Thunderstorm during incubation period, Flood during incubation

- DRG = 102, Admission

There is always a DRG, but particular storms may or may not be present, and a hospital admission (Admit) may or may not be present. Note that quantity is not considered here. The number of thunderstorms or the number of admits is not recorded, just the presence or absence of the "item".

We are interested in rules of the general form (DRG & Storm Type => Admit). Given the 23 levels of DRG and 5 storm types, there are 115 rules to be considered. The SAS EM Association Node can be configured with various properties in order to increase or decrease the number of rules generated for consideration. The figure below shows part of the output generated by SAS EM for our analysis.


Results - Node: Association Diagram: sas2011_v7

File Edit View Window

Table: Statistics Plot

Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1
0.58	5.22	0.30	8.94	2199.0	TS ==> W & F	TS	W & F	TS
5.84	52.21	0.30	8.94	2199.0	W & F ==> TS	W & F	TS	W
3.46	29.58	0.33	8.55	2412.0	WS & F ==> C	WS & F	C	WS
1.13	9.66	0.33	8.55	2412.0	C ==> WS & F	C	WS & F	C
17.97	99.01	0.33	5.51	2412.0	F & C ==> WS	F & C	WS	F
0.34	1.86	0.33	5.51	2412.0	WS ==> F & C	WS	F & C	WS
3.26	17.68	0.33	5.42	2412.0	WS & C ==> F	WS & C	F	WS
1.89	10.25	0.33	5.42	2412.0	F ==> WS & C	F	WS & C	F
3.26	15.76	0.01	4.83	99.00	C & 68 ==> F	C & 68	F	C
3.46	16.56	0.01	4.78	99.00	F & 68 ==> C	F & 68	C	F
5.84	27.62	0.03	4.73	219.00	F & 21 ==> TS	F & 21	TS	F
3.46	16.13	0.16	4.66	1142.0	admit & F ==> C	admit & F	C	admit
0.98	4.57	0.16	4.66	1142.0	C ==> admit & F	C	admit & F	C
3.26	14.42	0.01	4.42	61.00	C & 21 ==> F	C & 21	F	C
5.84	25.33	0.01	4.34	76.00	F & 423 ==> TS	F & 423	TS	F
0.24	1.02	0.03	4.27	239.00	F ==> C & 90	F	C & 90	F
3.26	13.92	0.03	4.27	239.00	C & 90 ==> F	C & 90	F	C
3.26	13.89	0.01	4.26	40.00	C & 69 ==> F	C & 69	F	C
3.46	14.71	0.09	4.25	619.00	WS & 99 ==> C	WS & 99	C	WS
0.58	2.48	0.09	4.25	619.00	C ==> WS & 99	C	WS & 99	C
1.37	5.69	0.20	4.16	1420.0	C ==> WS & 96	C	WS & 96	C
3.46	14.41	0.20	4.16	1420.0	WS & 96 ==> C	WS & 96	C	WS
5.84	24.19	0.20	4.14	1446.0	W & 21 ==> TS	W & 21	TS	W
0.83	3.43	0.20	4.14	1446.0	TS ==> W & 21	TS	W & 21	TS
3.46	13.88	0.03	4.01	239.00	F & 90 ==> C	F & 90	C	F

Most of the generated rules have very low support, which means that a particular DRG and storm type rarely accompany an Admit. 79 of the rules (69%) have a Support less than .01%.

The key association metric of interest for this analysis is Lift. A Lift greater than 1 suggests a positive correlation, while a Lift equal to 1 suggests no relationship. If Lift is less than 1, a negative association is implied. As we are only interested in the effect of storms increasing Admits associated with a DRG, rules with Lift less than or equal to 1 are set aside. Of the 115 rules generated, 27 (23%) have Lift > 1. The table below shows the key association metrics for the 10 rules with the highest Lift.

Rule	Count	Confidence (%)	Support (%)	Lift
C & 89 ==> admit	2709	60.85	0.38	2.3
WS & 89 ==> admit	14522	53.94	2.01	2.04
C & 98 ==> admit	1350	47.43	0.19	1.79
F & 89 ==> admit	1901	46.46	0.26	1.76
C & 79 ==> admit	956	43	0.13	1.63
WS & 98 ==> admit	5534	40.95	0.77	1.55
F & 98 ==> admit	900	40.09	0.12	1.52
WS & 79 ==> admit	4607	38.27	0.64	1.45
TS & 89 ==> admit	2640	35.11	0.37	1.33
W & 89 ==> admit	11890	34.47	1.65	1.3

Although a Lift greater than 1 suggests a positive association, SAS EM does not specify if the association discovered in the data is statistically significant. Significance can be tested via a traditional 2x2 cross tab analysis of counts. An example 2x2 table is shown below. The counts in the table are not part of EM results, but are calculated separately.

Rule: (WS & 89 => ADMIT)

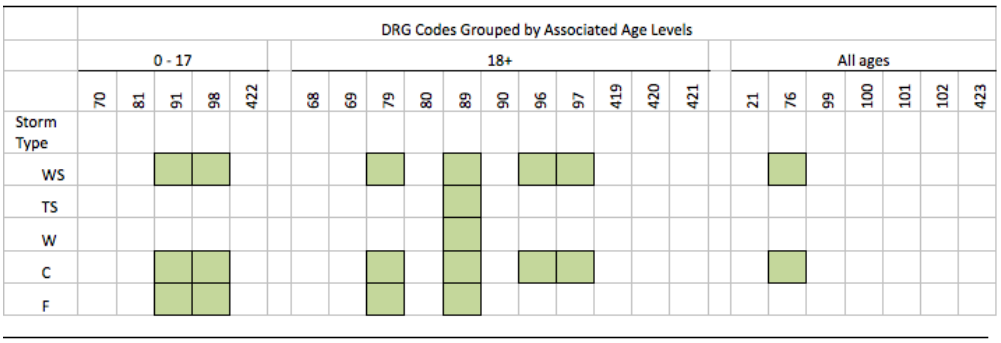
		"Admit" 1	"Admit" 0	
"89"	1	14522	12403	26925
	0	176241	518597	694838
		190763	531000	721763

Even though counts are not readily available, a Chi-Square statistic can be calculated via Support, Confidence, and Lift metrics (*Alvarez, 2003*). A Chi-Square statistic was generated for the 27 rules. Using a significance level of .01, 7 were found not to be statistically significant. For example, when "Cold" was matched with either DRG 70, 90, 420, or 422, the rule generated Lift > 1, but did not generate a significant Chi Square statistic.

This leaves 20 rules that are statistically significant. The table below shows the 3 rules with the highest and lowest Chi-Square values.

Rule	Lift	Chi Sq
WS & 89 ==> admit	2.04	10881.44
C & 89 ==> admit	2.3	2729.21
WS & 98 ==> admit	1.55	1493.74
WS & 76 ==> admit	1.11	24.88
F & 79 ==> admit	1.14	12.97
F & 91 ==> admit	1.11	6.77

There are very low and very high Chi-Square values. Although this is not a test of effect size, the extremely large Chi Square values are possibly due to large effects of the particular storm on the particular DRG related admits. The table below summarizes the pattern of results for the 20 statistically significant rules. DRG 89 is pneumonia. All storm types are related to jumps in admits for this already high-frequency DRG. Winter storms and Cold storms affect the most DRGs, while Thunderstorms and Wind only affect a single DRG.



Now that specific important relationships have been identified, this information can be used and/or investigated further during the Predictive Modeling phase of this investigation.

Analysis Task 2: Predicting Admission Counts

Our goal is a predictive model to predict the number of infectious diseases by DRG Group, area code, age group and week in the Score Dataset.

Our data mining task has been specified as a prediction problem rather than classification. Although the number of admits can only take positive integer values {0, 1, 2, ..} we decided to allow our predictions to take positive values on the interval scale. In other words, we are allowed to predict 1.3 admits.

When arriving at this decision we considered how the predictions would be used. The problem statement asks us to quantify the impact of weather storms on the use of health facilities. We were conscious that our predictions in the Score Dataset (at the DRG Group, Area code, age group and week level) would be assessed at the aggregate level. For example, the Institute for Improved Community Health may use our predictions to assess the impact on total DRG admits for a particular area code. They would achieve this by aggregating our predictions at the area code level. However if our predicted admits are constrained to be integer values, this introduces significant error into our modeling predictions when assessed at the aggregate level. More detail is provided as an Appendix.

Target Variable

Although our scored predictions will be for the number of admits, there are several ways in which we can set up our target variable to achieve this goal.

1. Binary Target

The distribution of admits is approximately binary- 98% of the distribution is either 0 or 1.

admits	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	531036	73.57	531036	73.57
1	177073	24.53	708109	98.10
2	10926	1.51	719035	99.62
3	1944	0.27	720979	99.89
4	545	0.08	721524	99.96
5	167	0.02	721691	99.98
6	64	0.01	721755	99.99
7	18	0.00	721773	100.00
8	15	0.00	721788	100.00
9	5	0.00	721793	100.00
10	4	0.00	721797	100.00
11	3	0.00	721800	100.00
12	3	0.00	721803	100.00
13	3	0.00	721806	100.00
14	2	0.00	721808	100.00

We create a new target variables called admits_binary, taking values:

0 when admits =0
1 when admits >0

Several predictive modeling algorithms exist for binary target data. They all can be configured to produce either a decision (0 or 1) or a posterior probability of success. As previously stated, ours is a prediction problem rather than a classification problem, and our predictions will be assessed at the aggregate level. We will therefore take the posterior probability estimate to be our prediction for the number of admits. To understand why this works, consider the expected number of admits for individual *i* :

$$\text{Expected \# Admits}_i = 0 \times P(\text{Admit} = 0 | X_i) + 1 \times P(\text{Admit} = 1 | X_i)$$
$$= P(\text{Admit} = 1 | X_i)$$

where X_i are the input variables relating to individual *i*

When we predict the total number of admits for several individuals i.e. all hospitals in an given area code, we sum together all the individual posterior probabilities of success.

Clearly then under this method the admit number predictions are constrained to be between 0 and 1. We can never predict more than one admit. However since the distribution of observed admits is approximately binomial, this method performs surprisingly well. Had the admit frequency of the population been much higher than this method would have deteriorated.

Instead of providing admits on an interval scale, we could have constrained predictions to be either 0 or 1 by selecting decisions rather than estimates. This method results in very poor predictive performance and was dropped. We refer the interested reader to the Appendix.

The following modeling algorithms were tested for the Binary Target

- Logistic Regression

- Decision Tree
- Neural Network

2. Count Target

The admit count takes integer values from 0 to 14. We can model this variable directly, however care should be taken to select the correct modeling methodology. Linear Regression requires the assumption that the model errors are independently and identically distributed. Since our data is essentially binary in nature, this assumption will not be satisfied. It is also a requirement that all admit predictions must be non-negative, and this could not be guaranteed with linear regression.

The following modeling algorithms were tested for the Count Target

- Decision Tree
- Neural Network

3. Admit Rate per Population

Common sense dictates that the admit count in an area code will have a strong positive association with its population. We can therefore attempt to predict the admit rate per population (admit rate) for each record in the scored data, and then convert this to a admit count prediction by multiplying through by population. Again care should be taken to select the appropriate modeling methodology

The following modeling algorithms were tested for the Admit Rate Target

- Neural Network
- Log Linear Models (Poisson, Negative Binomial, Zero Inflated Poisson)

Modeling Algorithms

Logistic Regression

We use Logistic Regression to model the binary target defined above. We predict

p - probability of a non-zero admit

by modeling the log odds as a linear combination of the input variables

$$\text{Log}(\, p\,/1\text{-}p) = B_0 + B_1X_1 + \ldots + B_nX_n$$

Where X_i are the input variables and B_i are the parameter estimates.

Decision Trees: Classification & Regression

Decision Trees are used to model both the binary and count targets. They are a rule based algorithm, that work by recursively splitting the data in the most homogeneous and predictive way. The advantage of Decision Trees is their ability to handle missing input data, which the other algorithms cannot do.

Neural Networks

Neural networks are applied to the binary, count and admit rate target variable.

A neural network, is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. It can be though of as a regression model on a set of derived inputs, called hidden units. In turn, the hidden units can be thought of as a regression on the original inputs. The hidden units "regression" include a default link/activation function, the hyperbolic tangent.

Neural networks have the ability to approximate virtually any continuous association between the inputs and the target. The analyst is required to specify the number of hidden units, which we do by trial and error. We also need to specify which inputs are to be used in the Neural Network, since there are no methods of variable section in the modeling process for a NN.

Although powerfully predictive, Neural Networks are difficult to explain which is a major drawback. We are not able to quantify the effect of storms on admits using this technique.

Log Linear Models

Poisson Regression

A variety of Log Linear models are available for modeling count data, the most famous being Poisson regression. The Poisson distribution has one parameter, the mean. A property of this distribution is that the variance is equal to the mean. In our context we us Poisson regression to model the admit rate, allowing different combinations of input variables to have different admit rates.

$$\text{Log (Count)} = \text{Log (Population)} + B_0 + B_1X_1 + \ldots + B_nX_n$$

$$\text{Log (Count)} - \text{Log (Population)} = B_0 + B_1X_1 + \ldots + B_nX_n$$

$$\text{Log (Count/Population)} = B_0 + B_1X_1 + \ldots + B_nX_n$$

$$\text{Log (Admit Rate)} = B_0 + B_1X_1 + \ldots + B_nX_n$$

where X_i are the input variables and B_i are the parameter estimates.

Other assumptions include

- The logarithm of the admit rate changes linearly with equal increment increases in the population variable.
- Changes in the admit rate from different model inputs are multiplicative
- Observations are independent

Log Linear models are not available under the standard Enterprise Miner 6 toolkit. To implement these types of models, we wrote Base SAS code to utilize the Genmod Procedure, which opens up a range of Generalized Linear models to us.

Negative Binomial

The Poisson distribution is often suggested for count data but found to be inadequate because the data displays far greater variance than that predicted by the Poisson. This is called overdispersion. It can shown that an alternative distribution called the Negative Binomial is an appropriate model when the over dispersion can be explained by heterogeneity of the mean over the population.

The Negative Binomial models admit rates in the same way as described for the Poisson.

Zero Inflated Poisson Model

In practice, we often see count data with excessive zero counts (no event), which may cause the deviation from the Poisson distribution - overdispersion or underdispersion. In this case the zero-inflated Poisson regression may be used, which is a two stage model.
One way to model this type of situation is to assume that the data come from a mixture of two populations, one where the counts is always zero, and another where the count has a Poisson distribution with mean μ. In this model zero counts can come from either population, while positive counts come only from the second one.
The distribution of the outcome can then be modeled in terms of two parameters, π the probability of 'always zero', and μ, the mean number of publications for those not in the 'always zero' group. A natural way to introduce covariates is to model the logit of the probability π of always zero and the log of the mean μ for those not in the always zero class.
Although estimated in a two stage process, the output from the Zero inflated Poisson model is an admit rate.

Input Variables

The following input variables were tested

	Variable Name	Description	Justification
Main Effects	**age_let**	Age Group	Socialization habits and immunity systems vary by age
	drg24	DRG code	DRG's have varying degrees of prevelance
	drg24cc	DRG Collapsed by Complication status	Dimensionality reduction
	ws_ind **w_ind** **c_ind** **ts_ind** **f_ind**	Windstorm Flag Wind Flag Cold Storm Flag Thunderstorm Flag Flood Flag	Captures the uplift in admit frequency during each storm incubation/association period
	qtr **month**	Quarter (1,2,3,4) month	DRG admits are seasonal in nature
	M_pop	Population imputation flag	Missing at random test
	Schooldays **workdays**		School/work provide social contact > increased disease transmission
	Weather variables	AvgLowT MinLowT AvgHighT MaxHighT prcp snow Daylight	Decision Tree only due to 40% Missing values
Interactions	**drg24cc x ws_ind** **drg24cc x w_ind** **drg24cc x c_ind** **drg24cc x ts_ind** **drg24cc x f_ind**		The uplift in admit frequency due to storm effects, may vary bewteen DRGs
	drg24cc x age_let		Allows the admit frequency of each DRG to vary by age
	age_let x qtr **age_let x month**		Certain age groups (65+) may be more prone to illness in winter months

To avoid redundancy, Quarter and month were not allowed in the same model. Similarly for DRG and DRG24cc. Weather variables were only tested in the Decision Tree models, as they handle missing values well. Other algorithms such as Logistic Regression and Neural Networks require missing value imputation, or the removal of records with missing values. Workdays was only used in the 18+ models. School days was only used in the under 18 models.

Model Assessment

Data was split into Training and Validation in the ratio 60:40. Models were trained on the training validation dataset, and assessed on their ability to correctly forecast admits in the validation dataset.

The forecast error is the difference between the actual value and the forecast value for the corresponding period.

$$E_t = Y_t - F_t$$

where E is the forecast error at period t, Y is the actual value at period t, and F is the forecast for period t.

We used the Average Square Error in the Validation data, as our model performance measure.

$$\text{ASE}= \frac{\sum_{t=1}^{N} E_t^2}{N}$$

Modeling Results

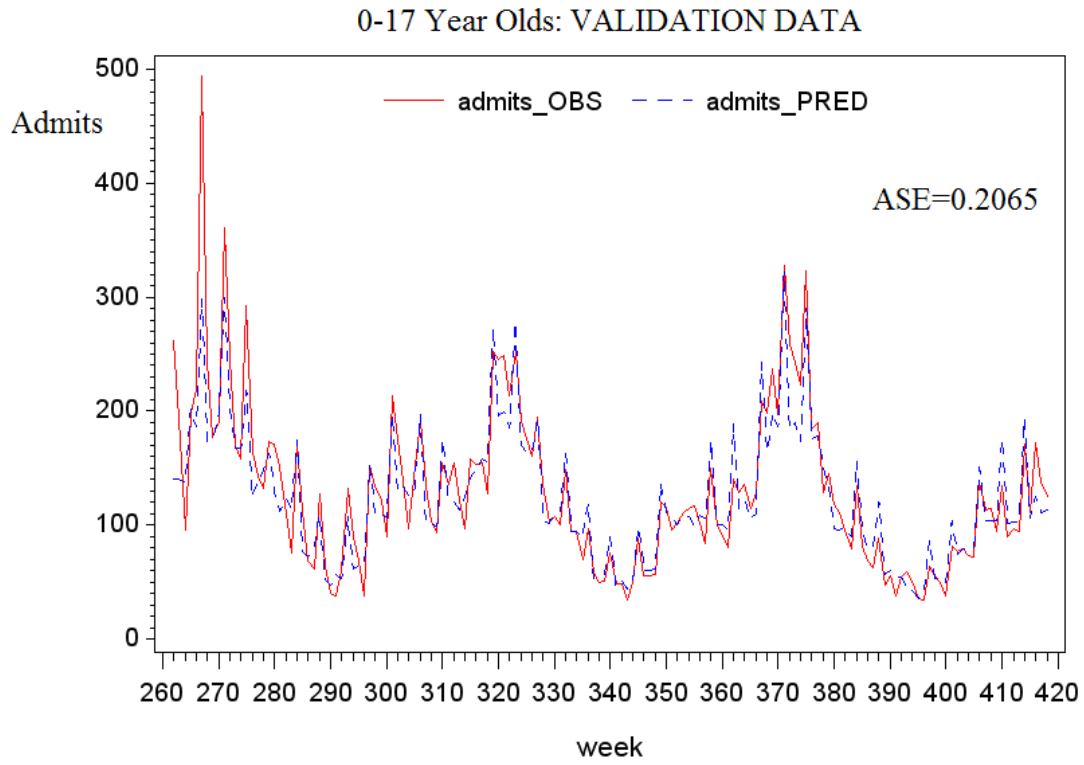
Certain DRG codes are only applicable to specific age ranges. At the early stages of modeling it became clear that two separate models were required for the different age ranges:

- Age Group 0-17
- Age Group 18-65+

Splitting the data in this way reduces the dimensionality of the problem, by reducing the number of interaction terms between Age group and DRG code. Using the insight derived at the data preparation stage it was felt that this was the most homogeneous way to split the data.

Age Group 0-17 Winning Model: Neural Network on Binary Target

The Neural Network trained on a Binary Target has the lowest ASE (0.2065) in the validation sample, and has therefore been selected as the winning model. This model used a step wise logistic regression model for variable selection. The performance in the validation data has been displayed below, with predicted and observed admits aggregated to the week level.

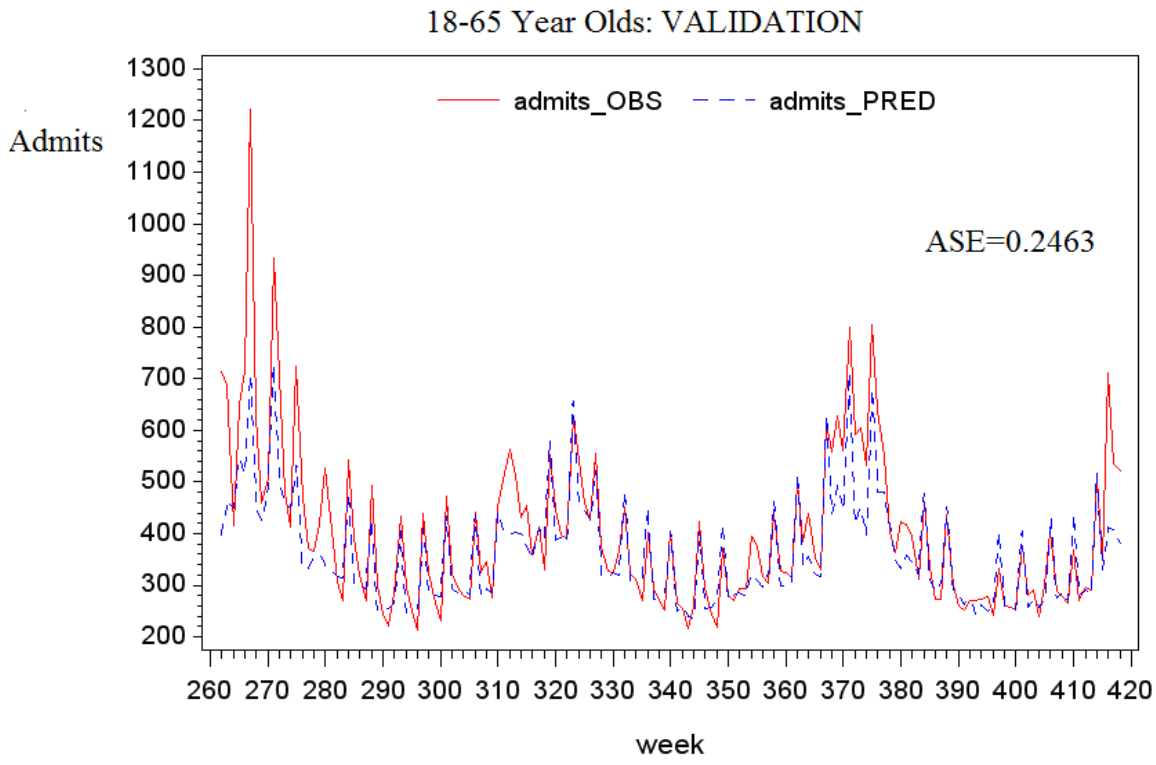


The performance of the other candidate models is provided for reference below.

MODEL	TARGET	ASE Training	ASE Validation
NN A (RegA step -)	Binary	0.208	**0.207**
DT Count	Count	0.209	0.208
DT C	Binary	0.213	0.211
ZIP	Rate	0.215	0.213
DT Weather	Binary	0.215	0.214
Reg A step -	Binary	0.216	0.214
Poisson	Rate	0.226	0.223
NN Count	Count	0.226	0.225
Negative Binomial	Count	0.284	0.281

Age Group 18-65+ Winning Model: Neural Network on Binary Target

The Neural Network trained on a Binary Target has the lowest ASE (0.2463) in the validation sample, and has therefore been selected as the winning model. This model used a step wise logistic regression model for variable selection. The performance in the validation data has been displayed below, with predicted and observed admits aggregated to the week level.



The performance of the other candidate models is provided for reference below.

MODEL	TARGET	ASE Training	ASE Validation
NN A (RegA step)	Binary	0.243	0.246
NN A (Reg A) (Keep 0 - 5)	Binary	0.234	0.246
DT Count (DRG Rate Collapse)	Count	0.243	0.247
NN Count	Count	0.245	0.247
DT Weather	Binary	0.245	0.249
DT Count (DRGcc)	Count	0.245	0.249
ZIP	Rate	0.255	0.258
Poisson	Rate	0.256	0.259
DT B	Binary	0.256	0.260
Reg A Bwd	Binary	0.258	0.261
Reg A BwdFull	Binary	0.258	0.261
Reg A Step	Binary	0.258	0.261
Reg A Fwd	Binary	0.258	0.261
Reg A Step (Keep 0 - 5)	Binary	0.249	0.261
Reg A Step (Keep 0 - 4)	Binary	0.243	0.261
Negative Binomial	Count	0.341	0.345

Predictive Modeling: Supporting Evidence

The winning model for both age groups is the Neural Network. Unfortunately, although Neural Networks provide excellent predictive power they provide very poor explanatory power. They do not produce rules or modeling parameters that we can easily interpret, which is their major drawback.

However for both age groups the Decision Tree model trained on the admit counts produced very strong candidate models. In contrast to Neural Networks, Decision Tree models are extremely easy to interpret. Enterprise Miner even offers the facility to produce "English Rules", providing a comprehensive set of rules that we can analyze. They also produce a Variable importance statistic, which can be quickly used to judge the predictive power of an input variable.

Variable Importance 0-17

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
log_pop		14	1	1	1
age_let		6	0.877836	0.898586	1.023638
NODE	Node	6	0.65098	0.651005	1.000039
Qtr		1	0.5923	0.687291	1.160377
ws_ind		4	0.40556	0.395066	0.974126
c_ind		3	0.377502	0.420373	1.113567
schooldays	Schooldays	4	0.222706	0.255936	1.149213
SchoolAge		1	0.070863	0.059581	0.840791
ts_ind		0	0	0	.
f_ind		0	0	0	.
w_ind		0	0	0	.
M_pop		0	0	0	.

Variable Importance 18-65+

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
log_pop		24	1	1	1
age_let		7	0.873096	0.912598	1.045244
NODE	Node	1	0.781922	0.795146	1.016911
ws_ind		3	0.686094	0.698305	1.017798
Qtr		4	0.294733	0.30096	1.02113
c_ind		4	0.128841	0.102569	0.796086
workdays	Workdays	2	0.036871	0.034741	0.942217
w_ind		2	0.029514	0.035293	1.195796
ts_ind		0	0	0	.
M_pop		0	0	0	.
f_ind		0	0	0	.

The Winter storms and Cold Storms have been identified as important in both Decision Tree models. The Thunderstorm Flood and Wind indicators are not important when it comes to predicting admit counts.

As well as the variable importance statistics, we can examine the individual rules. We take a weighted average of all the decision tree rules to arrive predictions for the average admits, in the presence/absence of storms.

AGES 0-17

	DRG	STORM	Number of Records	Average Admit Count
RULE 1	98, 99, 100, 102	c_ind EQUALS 0	6212	0.44
	98, 99, 100, 102	c_ind EQUALS 1	1004	0.74

	DRG	STORM	Number of Records	Average Admit Count
RULE 2	91, 422, 76,70, 81, 423, 21, 101	ws_ind EQUALS 0	55817	0.21
	91, 422, 76,70, 81, 423, 21, 101	ws_ind EQUALS 1	12192	0.28

AGES 18-65+

	DRG	STORM	Number of Records	Average Admit Count
RULE 1	Everything else	ws_ind EQUALS 0	146742	0.20
	Everything else	ws_ind EQUALS 1	33397	0.27

	DRG	STORM	Number of Records	Average Admit Count
RULE 2	89, 79,	ws_ind EQUALS 0	112534	0.35
	89, 79,	ws_ind EQUALS 1	23340	0.70

Conclusions

A key component of the association of infectious disease to weather is the belief that such a disease is much more likely to be spread in the community when people spend more time indoors in close contact with each other

Our analysis confirms this belief- the impact of Weather storms has been quantified for the various DRG hospital admits, using a variety of methods.

Winter Storms and Cold storms have the most impact, and we believe this is partly due to the extremely cold weather conditions they bring. Flood events have also been identified as important, and we believe that this is due to the high levels of rain fall associated with these storms. These associations have been verified using both Market Basket analysis and a variety of Predictive models. This lends strength to their validity.

A weak association was found between Wind Storms and Thunderstorms, but this association was not confirmed in the Predictive modelling approach.

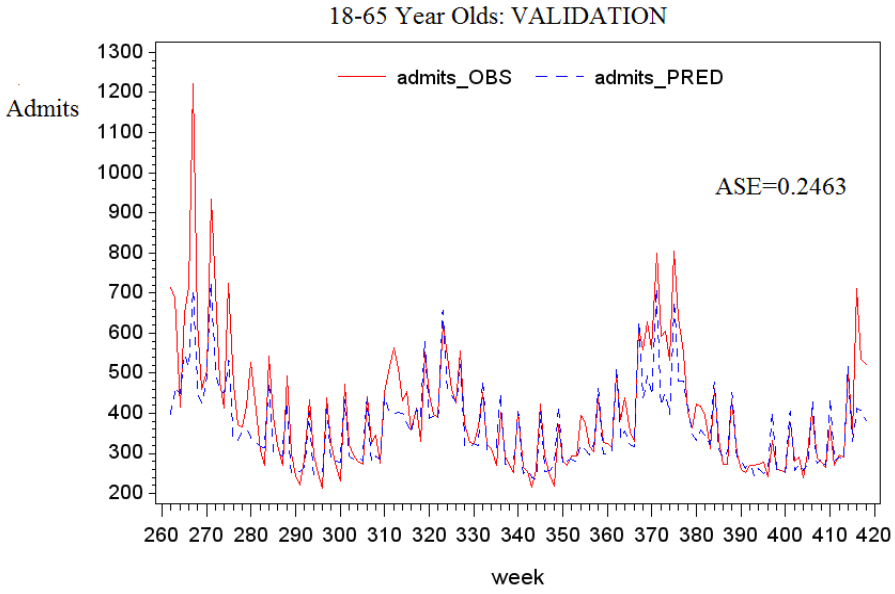
Future Improvements

The quality of the Weather data was poor- the high volumes of missing records forced us not to use this data as inputs to the predictive model. We feel that improving data capture would enable more accurate predictive models to be achieved.

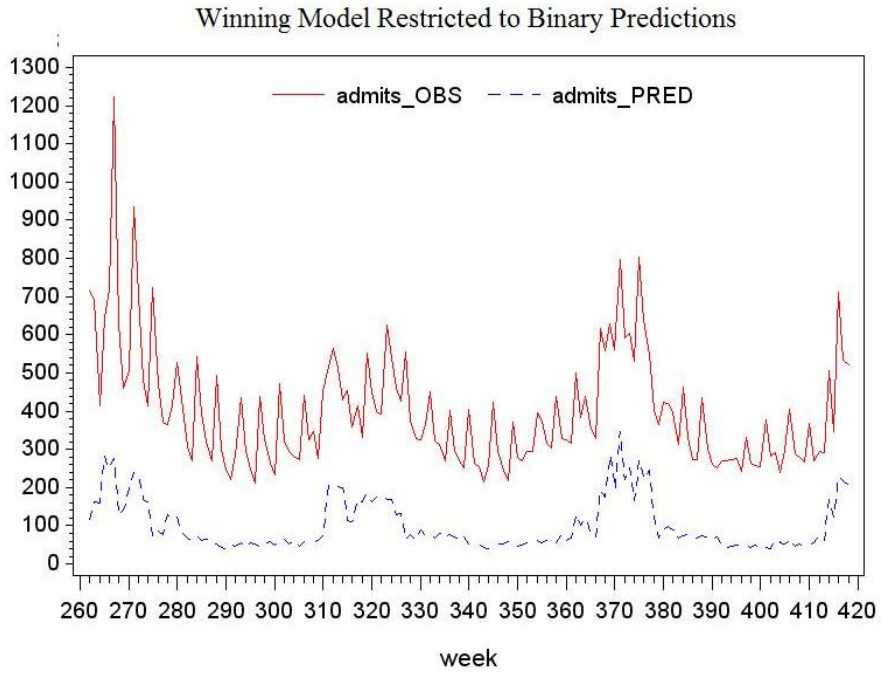
Other important considerations for spread of infectious disease are not part of this analysis. Sources suggest that key factors other than weather affect spread of infectious disease, such as Nutrition, Socio-Economic Status, Education, and Access to Health Care. Etc. Adding these to the model would provide a more complete understanding of the drivers behind hospital DRG admits.

Appendix

Winning Model for 18-65 year olds: Admit predictions were made on the continuous scale between 0 and 1, using the probability of admit_binary =1.



The Winning model shown above, but with admit predictions were decisions- predictions are either 0 or 1. This approach results in significant under prediction



References

World Health Organization 2005: Using Climate to Predict Infections disease epidemics http://www.who.int/globalchange/publications/infectdiseases/en/index.html

Halstead SB. (1996) Human factors in emerging infectious disease. WHO EMRO, 2: 21–29.

MacDonald G. (1957) The epidemiology and control of malaria. London, Oxford University Press

Mellor PS, Leake CJ. (2000) Climatic and geographic influences on arboviral infections and vectors. Revue Scientifique et Technique de l'Office International des Epizooties, 19:41–54.

Detinova TS. (1962) Age-grouping methods in Diptera of medical importance with special reference to some vectors of malaria. World Health Organization Monograph Series 47:13–191.

Alvarez, Sergio A. "Chi-squared computation for association rules: preliminary results," July, 2003, < http://www.cs.bc.edu/~alvarez/ChiSquare/chi2tr.pdf>.

Gottlober, Paul, ed. "Medicare Hospital Prospective Payment System: How DRG Rates Are Calculated and Updated," August, 2001.