Examining Salmonella Data from the CDC

Shoun Abraham
Marquette University
Milwaukee, WI

Quinn Furumo
Marquette University
Milwaukee, WI

Xavier Gomez
Marquette University
Milwaukee, WI

Nathan Lang
Marquette University
Milwaukee, WI

ABSTRACT

UPDATED—May 11, 2018.The aim of this paper is to propose a model for the prediction of the occurrence of hospitalization from Salmonella infection given certain criteria. We used the Center for Disease Control's National Outbreak Reporting System (NORS) dataset on Salmonella infection in the United States from 1998-2016. Although much of the data is categorical variables, we made these dummy variables to be able to use in a predictive model. Using these new variables, illnesses, and deaths we predicted the occurrence of hospitalization. We used a linear regression model with 10-fold cross validation. First, we selected the specific attributes based on significance level to run our model. Then, we also used recursive feature elimination (RFE) to select 10 attributes. We were able to predict with 83% accuracy if there would be an occurrence of hospitalization.

KEYWORDS

Salmonella; Salmonella Outbreak; Center for Disease Control; Contamination; Foodborne Illness; Preparation

	Illnesses	Hospitalizations	Deaths
count	19986.000000	16340.000000	16374.000000
mean	19.402982	0.952387	0.022047
std	48.622583	5.250625	0.380223
min	2.000000	0.000000	0.000000
25%	3.000000	0.000000	0.000000
50%	8.000000	0.000000	0.000000
75%	19.000000	1.000000	0.000000
max	1939.000000	308.000000	33.000000

Figure 1: Descriptive Stats

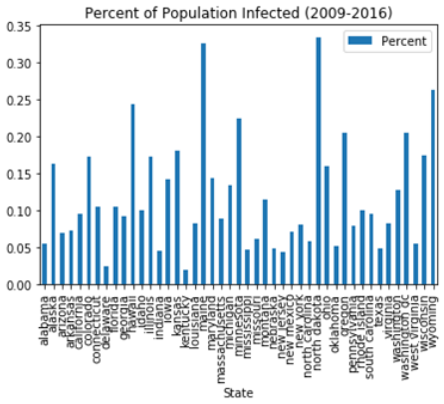

Figure 2: Percent of population infected per State

	Year	Month	State	Genus Species	Serotype or Genotype	Etiology Status	Location of Preparation	Illnesses	Hospitalizations	Deaths	Food Vehicle	Contaminated Ingredient
0	2009	1	Minnesota	Norovirus	NaN	Suspected	Restaurant - Sit-down dining	2	0.0	0.0	NaN	NaN
1	2009	1	Minnesota	Norovirus	NaN	Confirmed	NaN	16	0.0	0.0	NaN	NaN
2	2009	1	Minnesota	Norovirus	NaN	Suspected	Restaurant - Sit-down dining	5	0.0	0.0	NaN	NaN
3	2009	1	Minnesota	Norovirus	NaN	Confirmed	Restaurant - "Fast-food"(drive up service or p...	3	0.0	0.0	NaN	NaN
4	2009	1	Minnesota	Norovirus	NaN	Confirmed	Restaurant - other or unknown type	21	0.0	0.0	cookies	NaN

Figure 3: Head of the dataset

INTRODUCTION

This data set comes from the Center for Disease Control's National Outbreak Reporting System (NORS) on salmonella. The data set tracks each recorded instance of salmonella infection in the United States from 1998 to 2016. The dataset contains roughly 240,000 entries, 20,000 unique rows, and 11 columns relating to Salmonella. There was a large amount of missing data, primarily in the Serotype or Genotype, Food Vehicle, and Contaminated Ingredient columns. Since the data is qualitative and cannot be assumed, we were unable to fill in those missing data points, as doing so would be no better than random guessing. The data comes from the CDC or from doctors and hospitals, so we can conclude that the data is accurate and trustworthy. However, the data set also only includes reported cases of salmonella, so it is not a complete count of all cases of salmonella which occurred between 1998 and 2016 in the United States. The head of this dataset is displayed in Figure 3.

INITIAL OBSERVATIONS

The data set contains several different descriptive columns of interest including Genus (the specific type of bacteria under the Salmonella family), Contaminated Ingredient, Location of Preparation, and Food Vehicle. We conducted preliminary analysis of the food groups with the highest rate of infection from the data set. However, due to the extreme breadth and diversity of the data types as well as the large quantity of missing data in the aforementioned columns, we elected not to use these data in our analytical model.

Since the majority of our data is descriptive, we isolated the numerical columns of Illnesses, Hospitalizations, and Deaths (Figure 1). We chose Hospitalizations as our focal point of analysis as we lack the infection rate in a population data for Illness analysis and the mean count of the data in Deaths was too low to provide meaningful analysis.

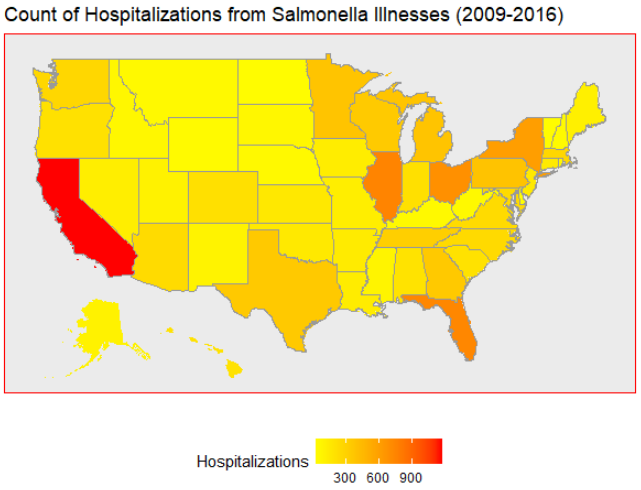

Figure 5: Choropleth of Hospitilizations from Salmonella Illness

(a) Intuitively, states with larger populations would have a larger number of cases of infection. This choropleth map follows this intuition as large population centers such as California, Illinois, Florida, and New York have high occurrences of Hospitalization, indicating more severe cases of Salmonella.

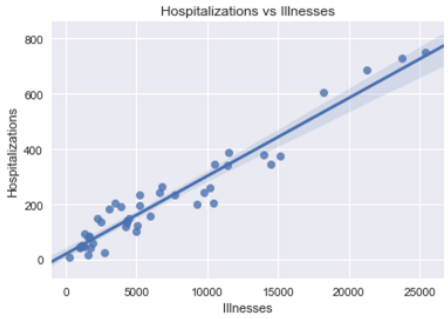

Figure 4: Hospitalizations vs Illnesses

INITIAL OBSERVATIONS CONT.

As Hospitalizations follow directly from Illnesses, the logical progression of analysis moved to finding correlations between Hospitalization count and Illness Count. In Figure 4, there is a noticeable linear relationship between Hospitalizations and Illnesses. From this initial correlation, we began to form models surrounding Hospitalizations, as it has the highest correlative potential of any of the numerical columns. A visualization about the percentage of the population infected with Salmonella (Figure 2) led us to create a chloropleth showcasing the number of Hospitalizations for each state (Figure 5). The four large population centers isolated from the choropleth map continue to have high values, while smaller states like Maine and North Dakota have high incidences of infection rate.

	Predicted: No	Predicted:Yes
Actual : No	TN = 4672	FP = 51
Actual : Yes	FN = 1164	TP = 109

Figure 7: Confusion Matrix (Initial Regression)

	Precision	Recall	F1-score	support
0	0.8	0.99	0.88	4723
1	0.68	0.09	0.15	1273
avg/total	0.78	0.8	0.73	5996

Figure 8: Confusion Matrix (Initial Regression) statistics

	Predicted: No	Predicted:Yes
Actual : No	TN = 4404	FP = 319
Actual : Yes	FN = 671	TP = 602

Figure 9: Confusion Matrix (Regression with RFE)

	Precision	Recall	F1-score	support
0	0.87	0.93	0.9	4723
1	0.65	0.47	0.55	1273
avg/total	0.82	0.83	0.82	5996

Figure 10: Confusion Matrix (Regression with RFE) statistics

DATA WRANGLING/CLEANING

Our data cleaning and data wrangling process took place over several steps. First, we took a look at the initial data set and its columns. The Location of Preparation column contained data points such as “Restaurant — Fast Food”, “Banquet Facility (food prepared and served on-site)”, and “School/college/university”. We removed the specific qualifiers like “Fast Food” from the Location of Preparation column and moved them to the new column “Specific Location” to remove clutter and allow for larger group analysis. Next, we looked at missing values. We did not fill the missing data within the descriptive columns as we cannot assume the data values for each specific situation. Similarly, we changed the missing values found in Hospitalizations and Deaths to 0 because we have no known values.

As we moved towards the graphing stage of our analysis, we created a choropleth map of Hospitalizations of each state from 2009-2016. We created a new dataframe by manually adding state location and population, while also removing multi-state and US territory cases.

Finally, for the implementation of Logistic Regression, we created dummy variables for all the categorical variables of our data, focusing specifically on predicting Hospitalization from the data. We also changed our Hospitalization column from a numeric column to a binary column. If there was one or more incidences of Hospitalization, we changed the numeric value to 1. We then removed the general parent variables of our dummy variables to allow for clean model creation, as well as those with large amounts of missing or unimportant data.

MODEL ANALYSIS

Logistic Regression was used to predict if Hospitalizations occurred given data specific to salmonella. Two rounds of Logistic Regression were used to obtain an optimal accuracy. In the first round of Logistic Regression, our group initially received singular matrix errors with sk-learn’s logistic regression. To resolve this issue, we removed the data from State, Genus, Serotype or Genotype, Etiology Status, Food Vehicle, and Contaminated Ingredient columns. We chose to do this to exclude unnecessary information.

We then ran logistic regression on the remaining data, resulting in p-values of less than 0.000 for Illnesses and Deaths and less than 0.000 for locations such as restaurants, religious institutions, prisons/jails, and farms. We then tested our model using a 70/30 train/test split and obtained an accuracy of 80%. In order to verify that our train/test split was not skewed, we also implemented a Ten-Fold Cross Validation on our model. The accuracy from that metric was 79%, indicating to us that our train/test split was valid. The statistics from our generated confusion matrix contained some problems. The f-1 score, a harmonic average between precision and recall, was only 0.15 for

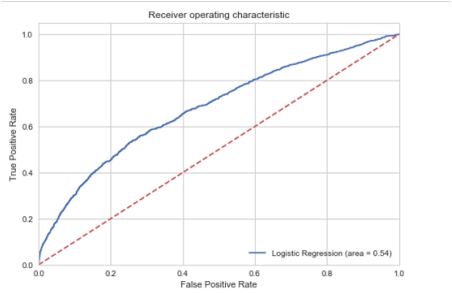

Figure 11: ROC curve: Initial Regression

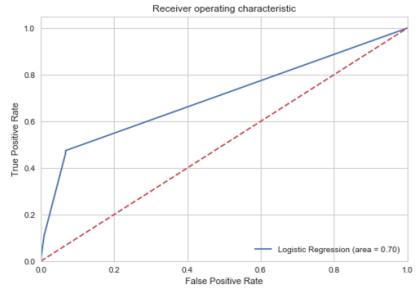

Figure 12: ROC curve using RFE

hospitalization. Also, we had over 1000 false negatives in our confusion matrix. A false negative in this context means that someone who should be hospitalized is not, resulting in adverse health effects. A visualization of the accuracy of our model is Figure 11 (ROC).

For our second round, we decided to use RFE (recursive feature elimination), an algorithm which automatically chooses the ten most statistically significant categorical variables from the data set. All ten values chosen were a specific Genus of Salmonella. From those ten values, seven proved to be statistically significant in the larger regression model. We applied the same train/test split and validation process as above, yielding 83% for both. The confusion matrix from this regression test yielded an f-1 score of .55 for hospitalization and had less than 700 false negatives, an improvement on the first regression test. The ROC curve was also better fit (see Figure 12), indicating a stronger model as a result of RFE analysis.

Some parts of the data set which potentially caused lower accuracy readings are the low amount of values >0 in the Hospitalizations column, the inability to fully add all portions of the data set (possibly skewing the overall correlative properties of the models), and innate connection between the values of Illnesses, Hospitalizations, and Deaths.

FINAL INFERENCES

Our strongest model for predicting the occurrence of Hospitalization given the data set has an accuracy of 83%. However, the false negative results in the confusion matrix are high, which is dangerous in relation to the real-world application of this model; if people have Salmonella and are not hospitalized within the confines of the model, then there will likely be large health consequences. However, regardless of how accurate any model created from this data set is, there are still a significant number of factors which go unaccounted for. The data set does not include data on the immune strength of the individuals infected, the time between infection, diagnosis, and treatment, the infection probability of the specific Salmonella bacteria, and a host of other case-specific values which are not included in this data set. This model does, however, give some more applicable insights into those most affected by Salmonella outbreaks. Our location data indicates that the locations which have the highest rate of infection are Restaurants, Caterers and other Private Institutions. While disease diagnosis and proper medical procedure (such as hospitalizing the patient through the use of our model) are important, preventative medicine is also essential in helping to stop infection in places where our data indicates.

REFERENCES

1. CDC. Multistate Outbreak of Human Salmonella Heidelberg Infections Linked to Ground Turkey (Final Update). Retrieved May 11, 2018 from https://www.cdc.gov/salmonella/2011/ground-turkey-11-10-2011.html

2. CDC. Multistate Outbreak of Human Salmonella Enteritidis Infections Associated with Shell Eggs (Final Update). Retrieved May 11, 2018 from https://www.cdc.gov/salmonella/2010/shell-eggs-12-2-10.html

3. Bishwa Adhikari, Frederick Angulo, Martin Meltzer. 2004. Economic Burden of Salmonella Infections in the United States. Retrieved May 11, 2018 from http://ageconsearch.umn.edu/bitstream/20050/1/sp04ad01.pdf

4. Giannella. 2004. Medical Microbiology. Retrieved May 11, 2018 from https://www.ncbi.nlm.nih.gov/books/NBK8435/

5. FDA. Foodborne Illnesses: What You Need to Know. Retrieved May 11, 2018 from https://www.fda.gov/Food/FoodborneIllnessContaminants/FoodborneIllnessesNeedToKnow/default.htm

6. Foodborne Illness. Salmonella. Retrieved May 11, 2018 from http://www.foodborneillness.com/salmonella_food_poisoning/

7. CIDRAP. USDA estimates E coli, Salmonella costs at $3.1 billion. Retrieved May 11, 2018 from http://www.cidrap.umn.edu/news-perspective/2010/05/usda-estimates-e-coli-salmonella-costs-31-billion

8. Angulo. 2006. Eating in Restaurants: A Risk Factor for Foodborne Disease? Retrieved May 11, 2018 from https://academic.oup.com/cid/article/43/10/1324/516737

9. Hitti. Salmonella: Frequently Asked Questions. Retrieved May 11, 2018 from https://www.webmd.com/food-recipes/food-poisoning/news/20080611/salmonella-frequently-asked-questions#1

10. NCBI. 2000. Salmonella Nomenclature. Retrieved May 11, 2018 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC86943

11. CDC. Multistate Outbreak of Human Salmonella I 4,[5],12:i:- Infections Linked to Alfalfa Sprouts (Final Update). Retrieved May 11, 2018 from https://www.cdc.gov/salmonella/2010/alfalfa-sprouts-2-10-11.html

12.Giannella RA. Salmonella. In: Baron S, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. Chapter 21. Retrieved May 11, 2018 from https://www.ncbi.nlm.nih.gov/books/NBK8435/