Enteric disease outbreaks in the US: Analysis of a dataset from the National Outbreak Reporting System

by

Brandon Meng

BS in General Biology, University of California, San Diego, 2016

Submitted to the Graduate Faculty of the

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Brandon Meng

It was defended on

June 16, 2021

and approved by

Committee Member: Jeanine Buchanich, PhD, MPH, MEd, Research Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Jeremy Martinson, D.Phil, Assistant Professor, Department of Infectious Diseases and Microbiology, Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Gong Tang, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Enteric disease outbreaks in the US: Analysis of a dataset from the National Outbreak Reporting System

Brandon Meng, MS

University of Pittsburgh, 2021

Abstract

With approximately 179 million cases occurring annually in the United States, acute gastroenteritis is a major public health issue. Cases are characterized by diarrhea and often followed by nausea, vomiting, fever, and abdominal pain. The Centers for Disease Control and Prevention tracks acute gastroenteritis outbreak data in the United States via the National Outbreak Reporting System (NORS). This thesis concerns the relationship between various factors of interest (year, season, region, setting, and etiology) and outcomes of illness, hospitalization, and death from person-to-person transmitted outbreaks of acute gastroenteritis in the United States. A relevant outbreak dataset was extracted from NORS. A negative binomial model was used to examine the various factors of interest on the number of illnesses and logistic regression models were used to examine the relationship between those factors and chance of hospitalization. Death in these outbreaks was compared with descriptive statistics due to sparsity. To account for missing values in setting and etiology, outbreaks with complete records and similar characteristics in other factors and number of illnesses were identified and the hot-deck method was used to impute missing values. Multiple imputation was used to summarize analysis results from datasets created with the hot-deck method. It was shown that that setting and etiology were by far the most influential factors on all three outcomes. Additionally, multiple imputation substantially reduced the variance estimates of some regression model parameters. Cases of acute gastroenteritis cause significant health and economic damage, so an examination of factors that are associated with

larger outbreaks is relevant to public health. Our results have important public health implications that mitigation of acute gastroenteritis outbreaks should be directed towards the school setting and Norovirus in particular. If policies are aimed at reducing severe outcomes, we should target *Salmonella*, *Clostridium*, and *Escherichia*, as these etiologies had the highest probabilities of hospitalization in this study.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I am grateful for all of the help and instruction I received from all of the faculty members from the Department of Biostatistics of Pitt Public Health.

In particular, I would like to thank Dr. Gong Tang for being my thesis advisor and guiding me through this entire process. I really appreciate all the instruction I have received on this project. This thesis would not have been possible without him.

I would also like to thank Dr. Jeanine Buchanich and Dr. Jeremy Martinson for agreeing to serve on my thesis committee. I am grateful for Dr. Buchanich being my academic advisor for so long and helped me through many hard times.

Thank you to Dr. Ada Youk for always making time to listen to my troubles and helping me sort things out.

I am thankful for my experience at Pitt and I appreciate everyone who contributed to it.

1.0 Introduction

With over 350 million annual cases in the United States, acute gastroenteritis (AGE) is a common cause of enteric illnesses leading to nausea, vomiting, diarrhea, and abdominal pain (Graves, 2013). This poses a global public health issue with childhood mortality in developing countries and significant economic burden in developed countries.

Most cases of acute gastroenteritis are characterized as viral infections or bacterial infections. Risk of infection is especially high in children due to their lack of immunity and lower likelihood to practice good hygiene habits. Rotavirus in particular is a very common cause in the younger demographic (Elliot, 2007), while Norovirus is more common in the older demographic (Chen 2017). Bacterial infections are also prevalent with common causes being species from *Escherichia*, *Salmonella*, and *Shigella*.

While most symptoms are usually mild and often require no treatment in developed countries (Wielgos, 2019), outbreaks still lead to significant health and economic impacts. A study examining the 5-year period of 2010-2014 in Belgium found that acute gastroenteritis caused 343 deaths, 27,707 hospitalizations, and 464,222 general practitioner consultations. The economic burden was estimated to represent direct costs of €112 million, indirect costs of €927 million, and an average total cost of €103 per case and €94 per person (Papadopoulos, 2019). As these figures show, public health initiatives directed towards mitigation of AGE are worth pursuing, even in developed countries like Belgium.

The CDC launched the National Outbreak Reporting System (NORS) in 2009 as an online platform for which health departments can enter outbreak information in the United States. Since the 1970s, national foodborne and waterborne disease outbreak surveillance have been core

functions of the CDC. The two surveillance systems handle this responsibility are the Waterborne Disease and Outbreak Surveillance System (1971-present) and the Foodborne Disease Outbreak Surveillance System (1973-present). While foodborne disease outbreak data have been collected electronically since 1998, NORS was designed to combine the outbreak reporting systems and improve national outbreak reporting with new components. Enteric disease outbreaks in the US are reported by local and state health departments to the NORS. The general flow of outbreak information to NORS usually consists of the following: 1) People are exposed to a pathogen; 2) People get sick and seek treatment; 3) Health department is notified of a possible outbreak; 4) Health department conducts an outbreak investigation; 5) Health department enters outbreak information into NORS; 6) CDC checks data for accuracy and analyzes; 7) Data are summarized and published (https://www.cdc.gov/nors/about.html).

NORS collects data on the following types of outbreaks in the United States: waterborne disease outbreaks, foodborne disease outbreaks, person-to-person transmitted disease outbreaks, animal contact disease outbreaks, environmental contamination outbreaks, and other enteric illness outbreaks. A study on the reporting period of 2009-2010 found that the primary reported mode of transmission in most AGE outbreaks was person-to-person at 52% (Hall, 2013). This trend has continued to recent times with person-to-person transmitted disease outbreaks at 63% of all enteric disease outbreaks reported to NORS from 2009-2018.

According to the NORS guidance documentation, the mode of transmission was defined as person-to-person if “the initial enteric illnesses were associated with direct contact with an infected person, their bodily fluids, or by contact with the local environment where the exposed person was simultaneously present with the infected person and may have had the opportunity for direct contact.” (NORS Guidance 7). Even though environmental contamination is often a factor

in person-to-person outbreaks, the primary mode of transmission is considered person-to-person if most of the patients had known direct contact or likely had the opportunity for direct contact with one another. For example, consider an outbreak that occurred in a long-term care facility. Multiple workers and residents become ill within a few days, with several more illnesses occurring over the next few weeks. Because multiple opportunities arose for direct contact among ill persons and there was evidence of propagated transmission, the mode of transmission was considered to be person-to-person.

A 2015 paper focused on the reporting period of 2009-2013 from the same NORS dataset at the time with outbreaks with the person-to-person, environmental, and unknown modes of transmission (Wikswo, 2015). This study mostly provided descriptive statistics and found that Norovirus was by far the most common etiology at 84% of outbreaks with observed etiologies. Outbreaks were found to be more frequent during the winter, with 53% occurring between December-February. Even though long term care facility was found to be the most common setting for all outbreaks, shigellosis and salmonellosis outbreaks were found to be especially prevalent in child care facilities. The paper noted the increase in reporting rates from 2009-2013 and stressed the importance of Norovirus as a very common pathogen from the dataset. The authors went on to highlight the utility of having a centralized database for more focused future studies on specific pathogens and their modes of transmission. They concluded that recommendations for prevention and control of AGE outbreaks through person-to-person contact, environmental contamination, and unknown modes of transmission depend primarily on appropriate hand hygiene, environmental disinfection, and isolation of ill persons.

The objective of this thesis is to examine the relationship between several factors of interest and three specific outcomes of interest (illness, hospitalization, and death) in enteric disease outbreaks using NORS data from 2009-2018.

The outcome of illness count was modeled with negative binomial regression. Hospitalization was modeled using grouped logistic regression models to examine and the association between demographic or etiologic factors and chance of hospitalization. Death was examined via descriptive statistics due to the scarcity of outbreaks with any death. While the intention is to examine how these same factors vary across different outcomes, specific variables will be removed from each model if they are found to not be useful to the model. This gives insight into risk factors of person-to-person transmitted enteric disease outbreaks and provide direction for where resources should be allocated for public health initiatives.

2.0 Background

2.1 A Dataset from NORS

Data were retrieved from NORS Dashboard, the CDC's webpage used to search and access the NORS data on enteric disease outbreaks. This site allows users to filter and extract data as well as visualize basic summary statistics on a few variables in each dataset.

This thesis was restricted to person-to-person transmitted enteric disease outbreaks from 2009 to 2018. This is because most prior research has focused on foodborne illness outbreaks, which have been recorded by the CDC since the 1970s, leaving research on person-to-person transmitted enteric disease outbreaks largely unexplored at this point in time. The dataset was also extracted using the single-state outbreaks filter, which excluded 13 multistate outbreaks, to ensure that there were only singular responses for the State variable. Each observation is an individual outbreak. Incidents were required to have at least two people contracting illnesses for them to be considered outbreaks and included in this dataset.

There were a total of 22,917 outbreaks from the extracted dataset. I found similar results as results as Wikswo in my analysis of the updated dataset and aim to examine the relationship between the outcomes and factors of interest from the dataset (Wikswo, 2015).

The extracted dataset included the following variables:

Table 1 Dataset Variables

Variable	Included or Excluded
Year	Included
Month	Included

State	Included
Primary Mode (of transmission)	Excluded
Etiology	Included
Serotype or Genotype	Excluded
Etiology Status	Excluded
Setting	Included
Illnesses	Included
Hospitalizations	Included
Info on Hospitalizations	Excluded
Deaths	Included
Info on Deahts	Excluded
Food Vehicle	Excluded
Food Contaminated Ingredient	Excluded
IFSAC Category	Excluded
Water Exposure	Excluded
Water Type	Excluded
Water Status	Excluded
Animal Type	Excluded
Animal Type Specify	Excluded

2.1.1 Outcomes of Interest

Out of the 21 total variables from the original dataset, three could be considered outcomes: "Illnesses," "Hospitalizations," and "Deaths." "Illnesses" was defined as "estimated total number of primary cases, including lab-confirmed and probable, based on the outbreak-specific definition." "Hospitalizations" was defined as "number of primary cases who were hospitalized." "Deaths" was defined as "number of primary cases who died." This means that the hospitalization counts and death counts are included in the illness counts. However, these three distinct outcomes will be examined separately.

2.1.2 Factors of Interest

Many of the 21 original variables are only relevant to different types of outbreaks not pertaining to this specific dataset. For example, variables like "Food Vehicle" and "Animal Type" are not relevant for this dataset which only contains person to person transmitted disease outbreaks. The dataset also contained variables that were clearly associated etiology, such as "Serotype or Genotype" and "Etiology Status". These variables are less relevant to the objectives of this thesis than the specific variables they correspond to (Etiology in this case) and could potentially cause multicollinearity issues with modeling and were not included in the analyses. After taking all these considerations, the finalized factors of interest from the dataset include: Year, Month, State, Setting, and Etiology.

2.2 Data Pre-Processing

Most factors were reclassified into fewer levels to examine trends and allow for more manageable regression analyses.

2.2.1 Year

The year variable was not reclassified from the original dataset and was treated as a discrete variable. This dataset contained year values 2009-2018 for person-to-person outbreaks.

2.2.2 Month

Historically, many person-to-person disease outbreaks have been shown to occur more frequently during winter months when weather is cooler. Despite the ubiquity of this trend, the exact mechanisms underlying these changes are still not well understood (Fares, 2013). The following factors have been proposed to explain the seasonality of various directly transmitted diseases:

1. Human Activity: With colder weather, people spend more of their time indoors, where pathogens thrive in crowded environments.
2. Pathogen Infectivity: Seasonal changes bring about changes in the physical environment like temperature, oxygen concentration, and humidity. For example, many micro-organisms like *E. coli* are much more stable in low-humidity conditions.
3. Immune System Function: Recent experimental studies on various animals including humans suggest that the immune system is weakened during the winter. Evidence

suggests that deficiencies in chemical compounds such as Vitamin D and Melatonin, both of which human bodies produce less of during the winter, can change normal immune system function.

The month variable was relabeled from 1-12 (January-December) to seasons as defined by northern meteorological seasons: Winter (December-February), Spring (March-May), Summer (June-August), and Autumn (September-November). As determining whether more outcomes occurred earlier or later in the year is not particularly helpful, months were reclassified to seasons, which hold more meaningful information by being related to climate trends.

2.2.3 State

The state variable from the original dataset (50 US States, Puerto Rico, and DC) was re-grouped into four US regions as defined by the US Census Bureau. This regrouping aims to reduce the number of levels for regression while also grouping together states that are likely to have similar climates based on similar geographic location.

2.2.4 Setting

The setting variable was re-grouped to combine similar levels together. This was done to identify potential trends across demographics usually found in specific locations. Long-term care facilities (LTCF), School, and Child Daycare had distinct enough demographics to remain unchanged for re-categorization. “Hospital” was combined with “Other healthcare facility” to create the new “Healthcare Facility” category. This categorization aims to examine outbreaks that occur in patients and healthcare workers. “Unknown” settings were converted and added to NA

for imputation. The remainder of the categories (all below 2.5%) were added together and used to create the new “Other” category.

2.2.5 Etiology

Etiologies from this dataset are either viral or bacterial. Norovirus is a common viral cause of acute gastroenteritis, while *Shigella* is a common bacterial agent.

In the United States, Norovirus is the most common cause of acute gastroenteritis across all age groups. Infection is characterized by diarrhea, vomiting, and stomach pain. Cases usually resolve themselves after 1-3 days. However, complications such as dehydration may occur, especially in the young, old, and those with pre-existing conditions such as diabetes. It can be transmitted directly from person to person or indirectly from contaminated water or food. Norovirus also is very contagious as fewer than twenty virus particles can cause an infection (Morillo, 2011).

Acute gastroenteritis can also be caused by bacterial infections from pathogens such as *Shigella*. *Shigella* is one of the leading bacterial causes of diarrhea worldwide and may cause dysentery, a type of gastroenteritis that results in diarrhea with blood, upon infection. Common symptoms are similar to those of Norovirus and usually last for several days. Cases usually resolve without specific treatment; however, complications can include reactive arthritis, sepsis, and seizures. As with most cases of bacterial infection, antibiotics can be administered to shorten the length of infection, but are usually only recommended for use in severe cases as resistance has become a common issue (Prince, 2010).

The original dataset contained 194 different etiologies, including viruses like Norovirus bacteria like *E. coli*. Since many of those etiologies are related to each other and can be aggregated, effort was made to categorize into a handful of pathogen groups.

Despite the clear difference between viruses and bacteria, both are classified relatively similarly using taxonomic systems. In both cases, genus refers to the classification above species and below family on the taxonomic hierarchy. In the original dataset, some entries for the etiology variable were genus (such as *Shigella*) while others were species (such as *E. coli*). Therefore, all etiology values were consolidated into viral genera or bacterial genera in this analysis for consistency.

Genera with too few outbreaks (<75 outbreaks observed) were collapsed into an "Other" category. Some outbreaks had multiple pathogens listed separated by semicolons such as "Adenovirus; Norovirus Genogroup II; *Clostridium* other." These outbreaks were collapsed into a "Multiple" etiology category. As almost all of the observations in this "Multiple" category contain Norovirus as one of the etiologies listed for this dataset, interpretation of this designation will be viewed as Norovirus in combination with at least one other etiology.

3.0 Statistical Methods

In this thesis, two well-established regression models were applied to study the association between factors of interest and outcomes in these outbreaks. Usually, the Poisson distribution is ideal for count data, but the assumption that the mean of the outcome is equal to the variance of the outcome is rarely the case in practice. In these situations, the negative binomial distribution is preferred to account for overdispersion. The negative binomial regression model was used to examine the relationship between the factors of interest and the number of illnesses in each outbreak.

Logistic regression models were used to study how the factors of interest related to the rate of hospitalization in each outbreak. Because there were missing values in etiology and setting, a hot-deck imputation algorithm was developed to impute missing etiology and setting by using the observed values from other outbreaks with similar characteristics in the dataset. The data were imputed several times and analysis results from the imputed datasets were summarized via multiple imputation (Rubin, 1987).

3.1 Negative Binomial Regression

The negative binomial is a conjugate mixture distribution for over-dispersed count data. It is a popular generalization of the Poisson distribution that relaxes the restriction of the variance equaling the mean. This is accomplished by modeling the Poisson heterogeneity with a gamma distribution.

Assume that: (1) given λ, Y follows a Poisson distribution with mean λ, and (2) λ follows a gamma distribution, F(k, μ), with probability density function

$$f(\lambda; k, \mu) = \frac{\left(\frac{k}{\mu}\right)^k}{(\Gamma(k))} \exp\left(-\frac{k\lambda}{\mu}\right) \lambda^{k-1}, \quad \lambda \geq 0$$

For this gamma distribution, we have

$$E(\lambda) = \mu, \quad var(\lambda) = \mu^2/k$$

Where $k > 0$ defines the shape. The degree of skewness decreases as k increases.

Marginally, the gamma mixture of the Poisson distributions is the negative binomial distribution for Y. The probability mass function of Y is

$$p(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1 - \frac{k}{\mu+k}\right)^y, \quad y = 0, 1, 2, \ldots.$$

In terms of the dispersion parameter $\gamma = 1/k$,

$$E(Y) = \mu, \quad var(Y) = \mu + \gamma\mu^2$$

The degree of overdispersion relative to the Poisson increases as γ increases. As γ approaches 0, the negative binomial distribution has var(Y) approaches μ and it converges to the Poisson distribution with mean μ.

Negative binomial regression model count data and allow μ to depend on covariates. Such models usually assume the same dispersion parameter γ for all observations (Agresti, 2013).

3.2 Logistic Regression

The logistic regression model is used to examine probability in a binary outcome. The model for $\pi(x) = P(Y = 1)$ at $x = (x_1, \ldots, x_p)$ for p predictors is

$$logit[\pi(x)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

An alternate formulation that directly specifies π(x) is

$$\pi(x) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}$$

We use indicator variables to denote categories for qualitative predictors. The parameter β_j is the effect of x_j on the log odds that Y = 1, adjusting for the other x_ks.

When more than one observation occurs at a fixed x_i, we record the number of observations as n_i and the number of successes as y_i. Then $\{Y_1, \ldots, Y_N\}$ are independent binomials with $E(Y_i) = n_i\pi(x_i)$, where $n_1 + \cdots + n_N = n$. Their joint probability mass function is proportional to the product of *N* binomial functions,

$$\prod_{i=1}^{N} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{n_i - y_i}$$

$$= \left\{\prod_{i=1}^{N} exp[\log(\frac{\pi(x_i)}{1 - \pi(x_i)})^{y_i}]\right\}\left\{\prod_{i=1}^{N}[1 - \pi(x_i)]^{n_i}\right\}$$

In the above likelihood function, we can model the data with $logit(\pi(x_i)) = \sum_j \beta_j x_{ij}$ (Agresti, 2013).

3.3 Cook's Distance

Cook's distance utilizes single-case deletion to calculate the influence of each observation on the fitted response values (Cook and Weisberg, 1982). Observations with Cook's distance greater than three times the mean Cook's distance can be considered outliers. As this metric was originally designed for linear models, the implementation for generalized linear models are

approximations, as described by Williams in 1987 (Williams, 1987). The formula for the Cook's distance metric applied to generalized linear models is the following:

$$\left(\frac{res}{1-hat}\right)^2 * \frac{hat}{dispersion * p}$$

Where *res* are Pearson residuals

hat is the hat matrix

p is the number of parameters in the model

and *dispersion* is the dispersion considered for the current model

3.4 Hot-Deck Imputation

Hot-deck imputation is an implicit and very popular imputation technique that identifies subjects with similar characteristics as the ones with missing values and impute the missing values by drawing observed values from those similar subjects with completely observed data. In this project, nearest neighbor imputation was used, meaning donors or observations with complete data in the close vicinity of each observation with missing values are identified and randomly allocated to impute each missing value. This method is convenient because it does not rely on explicit model fitting, and is less sensitive to model misspecification and violation of assumptions from methods like regression imputation (Andridge and Little 2010).

Contingency tables and chi-square tests of independence are usually used to examine whether a discrete variable to be imputed is associated with the other discrete variables. Regression analyses can be used to identify continuous variables that are related to discrete variables that are subject to missing values. In general, one can define a distance metric between any pair of

observations based on completely observed variables that are associated with variables that are subject to missing values. For each incomplete case, complete cases within its neighborhood of pre-determined size form the adjustment cells are potential donors. Let $x_i = (x_{i1}, \ldots x_{iq})$ be the values of q covariates from subject i and $C(x_i)$ denote the adjustment cell where subject i falls. Then one can randomly draw observed values from $C(x_i)$ to impute the missing values of x_i in the hot-deck imputation (Little and Rubin, 2002).

In practice, an adjustment cell based on discrete variables can be formed by matching those discrete variables exactly. In this thesis, etiology and setting are the two variables that are subject to missing values. It is noted that year, season, and region are the fully observed discrete variables that are highly associated with these two variables. The number of illnesses, as a continuous variable, is also highly associated with etiology and setting. We will consider creating adjustment cells based on these variables to impute missing values in etiology and setting.

3.5 Multiple Imputation

In general, missing values are imputed randomly via an explicit or an implicit model. Because of the randomness in the imputation, imputed values and the resulting parameter estimates will vary from imputation to imputation. Multiple imputation is a method of imputing missing values multiple times and making inferences by incorporating both within-imputation and between-imputation variation in parameter estimates derived from imputed datasets. In practice, we will repeat the same imputation algorithm multiple times. Then we will analyze each imputed dataset and summarize the analysis results from these multiply imputed datasets.

Analysis on single imputed data usually considers variation within the imputed data. Compared with single imputation, multiple imputation also considers variation due to imputation.

The analysis of multiply-imputed datasets is rather straightforward. Each imputed dataset is analyzed using the same analysis for complete rectangular data. Parameters are estimated using the following formula:

$$\overline{\theta_D} = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$$

Where $\hat{\theta}_d$ is the parameter estimate from a single imputation

D the number of imputations, and

$\overline{\theta_D}$ the average of parameter estimates from the *D* multiply imputed datasets

Because imputations are conditional draws rather than condition means, they provide valid estimates for a wide range of estimands, while averaging also increases efficiency. Variability has two components:

1. Average Within-Imputation Variance
    a. $\bar{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d$

        Where W_d is the within-imputation variance estimate from the *d*th imputed dataset
2. Between Imputation Variance
    a. $B_D = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \overline{\theta_D}\right)^2$

Total variability associated with θ_D is

$$T_D = \bar{W}_D + \frac{D+1}{D}B_D$$

Because the sample size is so large for this dataset, the standard normal distribution will be used as the reference distribution for interval estimates and significance tests.

4.0 Analysis Results

This section will be focused on analyses of the relationship between factors of interest and outcomes of illness, hospitalization, and death from the extracted dataset.

To address missing values from the setting and etiology variables, multiple imputation was used before fitting the same model used in the complete case analyses. Coefficient estimates were compared between the model fit on the data before and after the imputation of missing values to examine the effect of imputation.

4.1 Summary Statistics of Factors and Outcomes of Interest

This section shows descriptive statistics of re-categorized variables and examines their frequencies with regards to number of outbreaks.

4.1.1 Year

The year variable was not reclassified from the original dataset and was treated as a discrete variable. This dataset contained year values 2009-2018 for person-to-person outbreaks.
The frequencies of outbreak year are shown in Table 2.

Table 2 Frequencies of Outbreak Year

Year	Number of Outbreaks (Percentage)
2009	1308 (5.71%)

	2010	1781 (7.77%)
	2011	1892 (8.26%)
	2012	2554 (11.14%)
	2013	2555 (11.15%)
	2014	2490 (10.87%)
	2015	2682 (11.70%)
	2016	2709 (11.82%)
	2017	2549 (11.12%)
	2018	2397 (10.46%)

Table 2 shows that there was an increase in number of outbreaks each year from 2009 to 2012, when the number of outbreaks each year stabilizes at around 2500 (about 11% of the total).

4.1.2 Season

The month variable was re-labelled as season as defined by northern meteorological seasons. The frequencies of outbreak season are shown in Table 3.

Table 3 Frequencies of Outbreak Season

Season	Number of Outbreaks (Percentage)
Winter	12130 (52.93%)
Spring	6932 (30.25%)
Summer	1150 (5.02%)
Autumn	2705 (11.80%)

Table 3 shows that the majority of outbreaks occur during winter (52.93%) while very few occur during summer (5.02%).

4.1.3 Region

The state variable was re-labelled as region as defined by the US Census Bureau. The frequencies of outbreak region are shown in Table 4. Population sizes for each region from the 2020 US Census were included as a reference.

Table 4 Frequencies of Outbreak Region

Region	Number of Outbreaks (Percentage)	Population Size from 2020 US Census
Midwest	8096 (35.33%)	68,995,685
Northeast	6793 (29.64%)	57,609,148
South	5165 (22.54%)	126,266,107
West	2863 (12.49%)	78,588,572

Table 4 shows that the Midwest region of the US had the most outbreaks at 35.33% while the West region had the least amount of outbreaks at 12.49%.

4.1.4 Setting

The setting variable was re-grouped to combine similar levels together. The frequencies of re-categorized setting are shown in Table 5.

Table 5 Frequencies of Outbreak Setting

Setting	Number of Outbreaks (Percentage)
Long Term Care Facility	13316 (58.11%)
School/College/University	2156 (9.41%)
Child Daycare	1492 (6.51%)
Healthcare Facility	832 (3.63%)
Other	1672 (7.30%)
NA	3449 (15.05%)

Table 5 shows that the majority of the outbreaks were at long-term care facilities (58.11%) while fewer occurred at School (9.41%) and Daycare (6.51%). A sizable portion of the setting values is missing at 15.05%.

4.1.5 Etiology

The etiology variable was condensed into bacterial genera and viral genera for consistency. The frequencies of re-grouped etiology are shown in Table 6.

Table 6 Frequencies of Outbreak Etiology

Etiology	Number of Outbreaks (Percentage)
Norovirus	15530 (67.77%)
Shigella	854 (3.73%)
Salmonella	219 (0.96%)
Clostridium	164 (0.72%)
Escherichia	138 (0.60%)

	Rotavirus	104 (0.45%)
	Sapovirus	78 (0.34%)
	Other	182 (0.79%)
	Multiple	611 (2.67%)
	NA	5037 (21.98%)

Table 6 shows that the majority of the outbreaks were caused by Norovirus at 67.77%. *Shigella* is noted as being 3.73% of all outbreaks while the rest of the single pathogen etiologies are all below 1%: *Salmonella* at 0.96%, *Clostridium* at 0.72%, *Escherichia* at 0.60%, Rotavirus at 0.45%, and Sapovirus at 0.34%. A large portion of the etiology values is missing at 21.98%.

4.1.6 Outcomes

The outcomes of interest from the dataset are counts of illnesses, hospitalizations, and deaths. Table 7 shows summary statistics of these three different outcomes.

Table 7 Outcome Summary Statistics

Variable	Mean (SD)	Median (1^{st} Quartile, 3^{rd} Quartile)	Missing Values (Percentage)
Illnesses	32.71 (42.20)	23.00 (12.00, 41.00)	0 (0%)
Hospitalizations	0.5342 (1.6988)	0 (0, 0)	2647 (11.55%)
Deaths	0.0452 (0.2803)	0 (0, 0)	2625 (11.45%)

By comparing the medians to the means, we can see from Table 7 that all three outcomes are heavily right-skewed. The data show that hospitalizations and deaths consist almost entirely of zero values as medians, 1st quartiles, and 3rd quartiles are all 0.

Illnesses does not have any missing values while hospitalizations and deaths have very similar amounts of missing values. Almost all of the observations with these missing values are missing values from both hospitalizations and deaths.

4.2 Preliminary Analyses

Complete-case analyses were conducted before imputation of missing values to examine the effects of the factors of interest on counts of illness, probabilities of hospitalization, and probabilities of death from outbreaks.

4.2.1 Illnesses

4.2.1.1 Model Fitting

The negative binomial model was chosen for illnesses because the outcomes are counts and the summary statistics of illness count from Table 7 showed a heavy right skew, meaning the data were overdispersed.

To ensure that the negative binomial model was appropriate, univariate negative binomial regressions with each covariate were used for number of illnesses to examine dispersion parameters for overdispersion.

Table 8 Estimated Dispersion Parameter from Univariate Negative Binomial Regression Models on Illness Count

Covariate		Dispersion Parameter
	Year	1.3901
	Season	1.3898
	Region	1.3775
	Setting	1.5320
	Etiology	1.4785

Table 8 shows that all of the univariate regression dispersion indices are significantly greater than 1, suggesting overdispersion is present, meaning that the negative binomial model is an appropriate fit.

To determine if all covariates were necessary for the model, likelihood ratio tests were run. Each likelihood ratio test compared the full model with all of the covariates against sub-models with each of the covariates removed.

Table 9 Likelihood Ratio Test against Sub-Models

Covariates		Test Statistic	P value
	Year	514.54	< 0.0001
	Season	91.28	< 0.0001
	Region	39.14	< 0.0001
	Setting	25424.59	< 0.0001
	Etiology	37184.86	< 0.0001

Table 9 shows that all of the covariates appear to be statistically significant predictors of the number of illnesses in the multivariate negative binomial regression model.

Residuals were plotted as diagnostics for how well the model fit the data. A quantile-quantile plot was used to examine how close the model residuals were to a normal distribution in Figure 1.

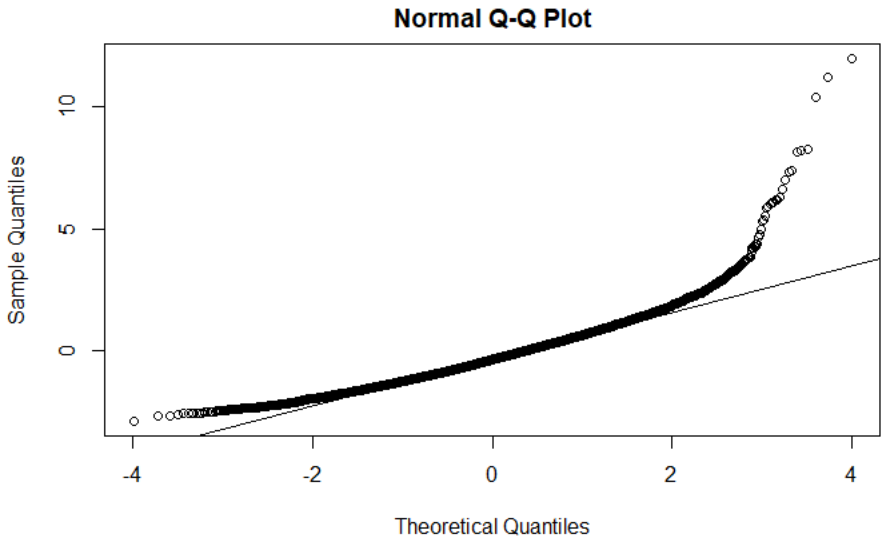

Figure 1 Quantile-Quantile Plot for Illness Count Model Residuals

The data points from this quantile-quantile plot mostly follow the line, showing that the distribution of the residuals mostly follows a normal distribution.

A histogram was used to examine the distribution of residuals for outliers and skewness in Figure 2.

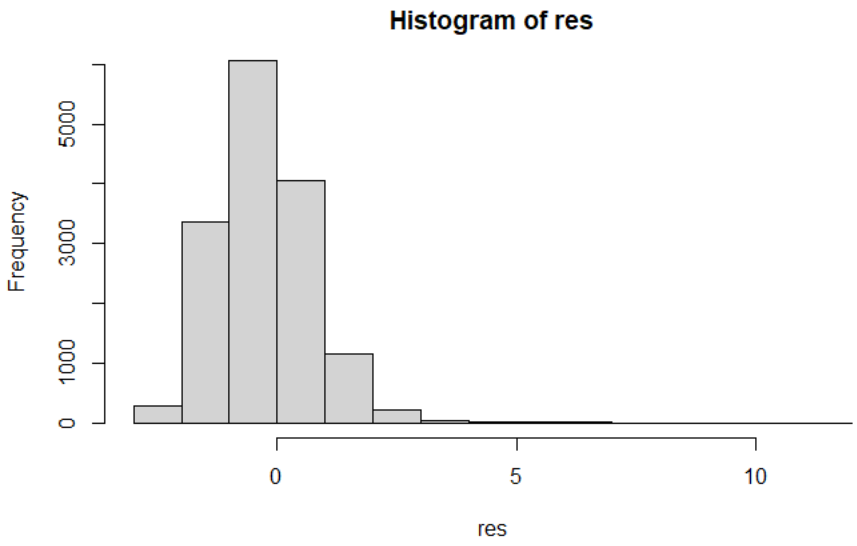

Figure 2 Histogram of Illness Count Model Residual Distribution

This histogram shows slight right skew, meaning that there are likely a couple observations with very high residual values.

The metric used to determine outliers from the model was Cook's distance. Cook's distance values were calculated for each of the 15,205 complete-case observations. Figure 3 shows which observations had the highest Cook's distance values.

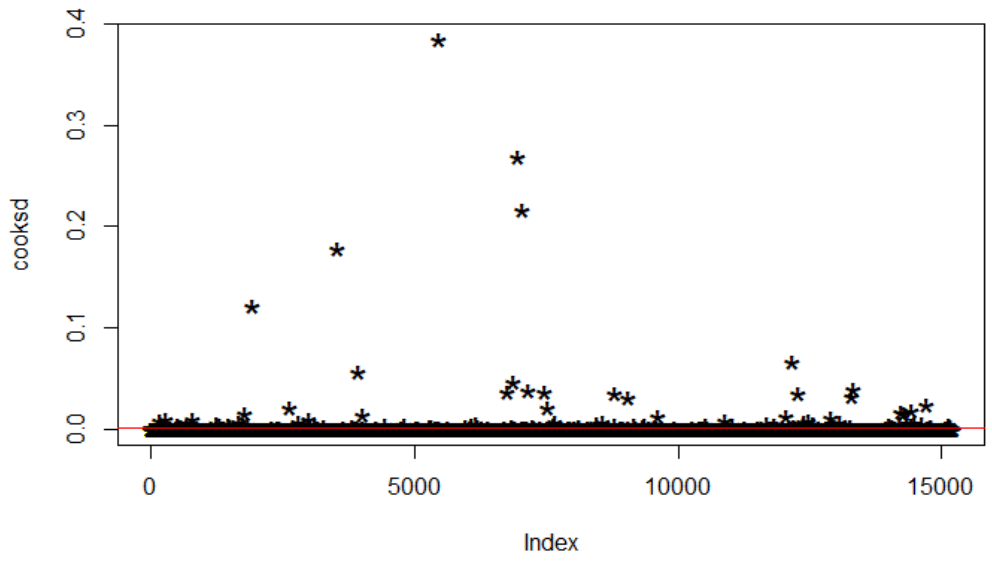

Figure 3 Outliers of Illness Count Model by Cook's Distance

Any observations with Cook's distance values of greater than four times the mean (0.000784 represented by the red line) were removed from the dataset to allow for a better model fit. The dataset with complete data observations was reduced from 15,205 outbreaks to 14,959 outbreaks after removal of outliers.

4.2.1.2 Examination of Outliers from Illness Model

Outbreaks considered outliers using the Cook's distance metric were separated into groups based on illness count quartiles.

Table 10 Outlier Outbreaks Sorted by Illness Counts

Illness Counts	Number of Outbreaks	Percentage
2-12	72	29.27%
13-23	50	20.33%

24-41	58	23.58%
42-180	66	26.83%
Total	246	

Table 10 shows that there was not a disproportionate amount of high outliers or low outliers with regards to illness count. Out of the total 246 outbreaks outliers using the Cook's distance metric, 66 outbreaks had illness counts greater than the 3rd quartile (41). These outbreaks with high illness counts were examined by setting and etiology, the two covariates with highest dispersion parameters and lowest likelihood ratio test p-values. This was done to examine potential trends in outbreaks that had higher than predicted illness counts.

Table 11 High Illness Count Outliers by Setting

Setting	Number of Outbreaks	Percentage
Long Term Care Facility	52	78.79%
School/College/University	9	13.64%
Child Daycare	1	1.52%
Healthcare Facility	2	3.03%
Other	2	3.03%
Total	66	

Table 11 shows that a majority of the high illness count outliers came from outbreaks in long term care facilities. This is expected as most of the outbreaks in the dataset were also in long term care facilities.

Table 12 High Illness Count Outliers by Etiology

Etiology	Number of Outbreaks	Percentage
Norovirus	61	92.42%
Shigella	1	1.52%
Rotavirus	1	1.52%
Multiple	3	4.55%
Total	66	

Table 12 shows that almost all of the high illness count outliers came from Norovirus outbreaks. This is expected as most of the outbreaks in the dataset were due to Norovirus.

4.2.1.3 Model Estimates

After removing the 246 outliers, a main effects negative binomial regression model was run with year, season, region, setting, and etiology as the covariates and the number of illnesses the count outcome. The estimates are presented in Table 13, where the baseline references for the covariates are:

1. Year: 2009
2. Season: Winter
3. Region: Midwest
4. Setting: Long Term Care Facility
5. Etiology: Norovirus

Table 13 Multivariate Negative Binomial Regression Model on Illness Count

Covariate		Multiplication Factor [95% CI]	P-value
Year			
	2009	1.00	
	2010	1.04 [0.96, 1.13]	0.3298
	2011	1.07 [0.99, 1.16]	0.0972
	2012	0.94 [0.87, 1.01]	0.0949
	2013	0.82 [0.77, 0.89]	< 0.0001
	2014	0.88 [0.82, 0.95]	0.0012
	2015	0.84 [0.78, 0.90]	< 0.0001
	2016	0.73 [0.68, 0.77]	< 0.0001
	2017	0.71 [0.67, 0.77]	< 0.0001
	2018	0.65 [0.61, 0.70]	< 0.0001
Season			
	Winter	1.00	
	Spring	0.88 [0.86, 0.91]	< 0.0001
	Summer	0.81 [0.76, 0.86]	< 0.0001
	Autumn	0.92 [0.88, 0.96]	0.0002
Region			
	Midwest	1.00	
	Northeast	1.11 [1.07, 1.15]	< 0.0001
	South	1.06 [1.03, 1.09]	0.0005

	West	1.07 [1.03, 1.12]	0.0009
Setting			
	Long Term Care Facility	1.00	
	School/College/University	1.84 [1.77, 1.92]	< 0.0001
	Child Daycare	0.64 [0.60, 0.68]	< 0.0001
	Healthcare Facility	0.70 [0.66, 0.75]	< 0.0001
	Other	1.04 [0.99, 1.10]	0.0744
Etiology			
	Norovirus	1.00	
	Shigella	0.68 [0.63, 0.73]	< 0.0001
	Salmonella	0.25 [0.21, 0.29]	< 0.0001
	Clostridium	0.39 [0.32, 0.48]	< 0.0001
	Escherichia	0.32 [0.28, 0.38]	< 0.0001
	Rotavirus	0.67 [0.57, 0.80]	< 0.0001
	Sapovirus	0.86 [0.72, 1.03]	0.0986
	Other	0.25 [0.22, 0.29]	< 0.0001
	Multiple	1.35 [1.26, 1.44]	< 0.0001

The coefficient estimates from Table 13 are displayed with a dot and whisker plot in Figure 4.

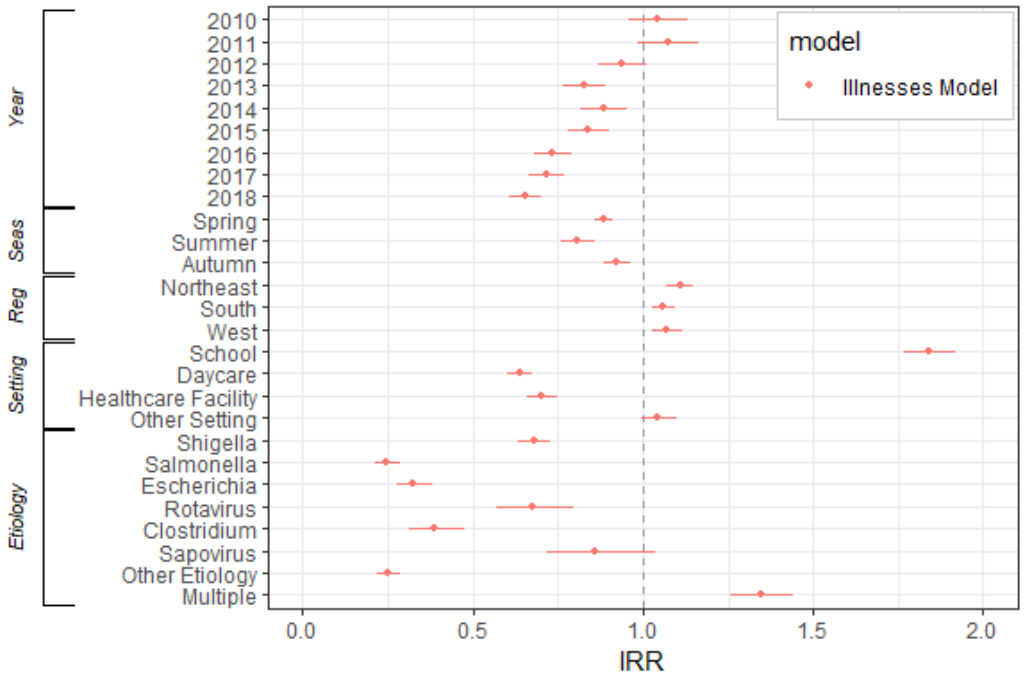

Figure 4 Multivariate Negative Binomial Regression Model for Illness Count

4.2.1.4 Interpretation of Model Estimates

In Table 13 and Figure 4, the multiplication factors provide an idea on how the average number of illness of an outbreak differentiates across different levels of a covariate. For example, on average, the number of illnesses per outbreak in 2010 was 1.04 times that of an outbreak in 2009. 2011 also shows a similar increase with a multiplication factor of 1.07. From 2011 to 2018, the multiplication factors show a decreasing trend with 2011 at 1.07, 2014 at 0.88, and 2018 at 0.65. This means that there are usually fewer counts of expected illnesses year to year relative to 2008 when holding the other factors constant.

All season multiplication factors are less than 1.0, meaning winter had the most expected illnesses per outbreak while summer (0.81) had the least. This means that on average, the number of illnesses from an outbreak in summer is 0.81 times that of winter. Spring and autumn have

similar numbers of illnesses during outbreaks compared to winter with multiplication factors of 0.88 and 0.92.

Region multiplication factors are all greater than 1, meaning on average, the Midwest has the lowest number of illnesses. The other regions all have slightly more numbers of illnesses compared to the Midwest, with Northeast at 1.11 times, South at 1.06 times, and West at 1.07 times.

School has a noticeably high multiplication factor at 1.84, meaning that on average, the number of illnesses at schools is 1.84 times that of long term care facilities. Daycare (0.64) and healthcare facility 0.70) appear to have significantly lower numbers of illnesses compared to LTCF.

Multiple (1.35) is the only etiology with a multiplication factor greater than 1.0. As all the other etiologies have factors below 1.0 by a rather noticeable margin, this means that, on average, Norovirus outbreaks have the highest numbers of illnesses out of the single pathogen etiologies. *Salmonella* (0.25), *Escherichia* (0.32), and *Clostridium* (0.39) have similarly low factors, meaning outbreaks with these etiologies usually have much lower numbers of illnesses compared to Norovirus outbreaks. Rotavirus [0.57, 0.80] and Sapovirus [0.72, 1.03] have especially wide confidence intervals, likely due to the scarcity of outbreaks with these either of these two as the etiology listed.

Overall, setting and etiology appear to have the most influence on illnesses per outbreak out of the factors of interest.

4.2.2 Hospitalizations

4.2.2.1 Model Fitting

For each outbreak, the number of hospitalized is part of the number of illnesses. Therefore, the binary outcome for this logistic regression model is whether or the person affected by the outbreak was hospitalized or not hospitalized.

To determine if all covariates were necessary to the model, likelihood ratio tests were run. Each likelihood ratio test compared the full model with all of the covariates against sub-models with each of the covariates removed.

Table 14 Likelihood Ratio Test against Sub-Models

Covariates		Test Statistic	P value
	Year	57.91	< 0.001
	Season	20.19	< 0.001
	Region	162.61	< 0.001
	Setting	4252	< 0.001
	Etiology	1481.5	< 0.001

All of the covariates appear to be statistically significant predictors of hospitalizations in the logistic regression model.

Various residual plots were examined in order to determine how well the model fit the data. A quantile-quantile plot was used to examine how close the model residuals were to a normal distribution in Figure 5.

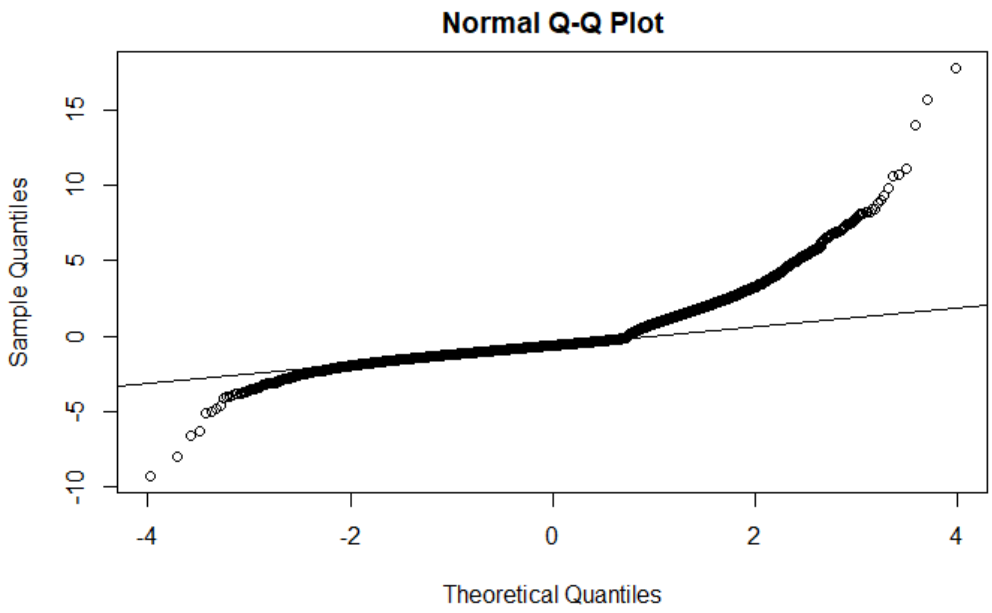

Figure 5 Quantile-Quantile Plot Hospitalization Model Residuals

This Q-Q plot shows that the residual distribution does not appear to resemble a normal distribution, meaning the model likely does not fit the data well. This suggests that the included variables are not adequate to predict the chance of hospitalization.

A histogram was used to examine the distribution of residuals for outliers and skewness in Figure 6.

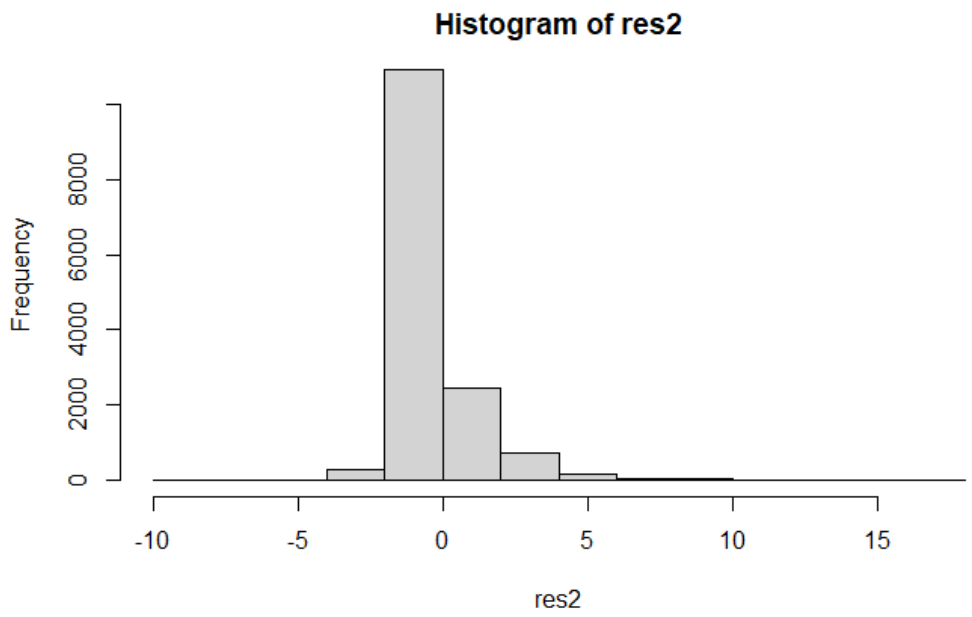

Figure 6 Histogram of Hospitalization Model Residual Distribution

This histogram shows that most of the residuals are slightly negative with more outliers on the positive tail end.

Because the hospitalizations variable has so few non-zero counts (only 26% of the observations with non-missing hospitalization values), removal of outliers actually increased model variance metrics; therefore, the model will be fit on the observations with complete-case data without removal of outliers.

4.2.2.2 Model Estimates

A Logistic Regression was run with Year, Season, Region, Setting, and Etiology as the covariates and Hospitalizations versus Not Hospitalizations as the dichotomous outcomes. The baseline references for the covariates are:

1. Year: 2009

2. Season: Winter
3. Region: Midwest
4. Setting: Long Term Care Facility
5. Etiology: Norovirus

Table 15 Multivariate Logistic Regression Model on Hospitalizations

Covariate		Multiplication Factor [95% CI]	P-value
Year			
	2009	1.00	
	2010	0.90 [0.79, 1.02]	0.0949
	2011	0.95 [0.84, 1.08]	0.4394
	2012	1.05 [0.94, 1.18]	0.3863
	2013	1.07 [0.95, 1.20]	0.2489
	2014	1.00 [0.89, 1.12]	0.9898
	2015	0.96 [0.86, 1.08]	0.5220
	2016	1.16 [1.04, 1.30]	0.0120
	2017	1.16 [1.04, 1.30]	0.0106
	2018	1.16 [1.03, 1.30]	0.0127
Season			
	Winter	1.00	
	Spring	1.01 [0.96, 1.07]	0.5901
	Summer	1.23 [1.10, 1.38]	0.0003
	Autumn	0.92 [0.86, 1.00]	0.0415
Region			

Midwest	1.00	
Northeast	0.86 [0.81, 0.91]	< 0.0001
South	1.02 [0.97, 1.08]	0.4311
West	1.39 [1.30, 1.49]	< 0.0001
Setting		
Long Term Care Facility	1.00	
School/College/University	0.07 [0.06, 0.08]	< 0.0001
Child Daycare	0.21 [0.18, 0.24]	< 0.0001
Healthcare Facility	4.05 [3.78, 4.34]	< 0.0001
Other	0.51 [0.46, 0.56]	< 0.0001
Etiology		
Norovirus	1.00	
Shigella	5.86 [5.14, 6.68]	< 0.0001
Salmonella	9.64 [7.54, 12.18]	< 0.0001
Clostridium	10.10 [8.31, 12.21]	< 0.0001
Escherichia	19.40 [15.23, 24.48]	< 0.0001
Rotavirus	2.96 [2.35, 3.67]	< 0.0001
Sapovirus	0.44 [0.23, 0.74]	0.0044
Other	4.96 [3.48, 6.88]	< 0.0001
Multiple	1.36 [1.23, 1.49]	< 0.0001

The coefficient estimates from Table 15 are displayed with a dot and whisker plot in Figure 7.

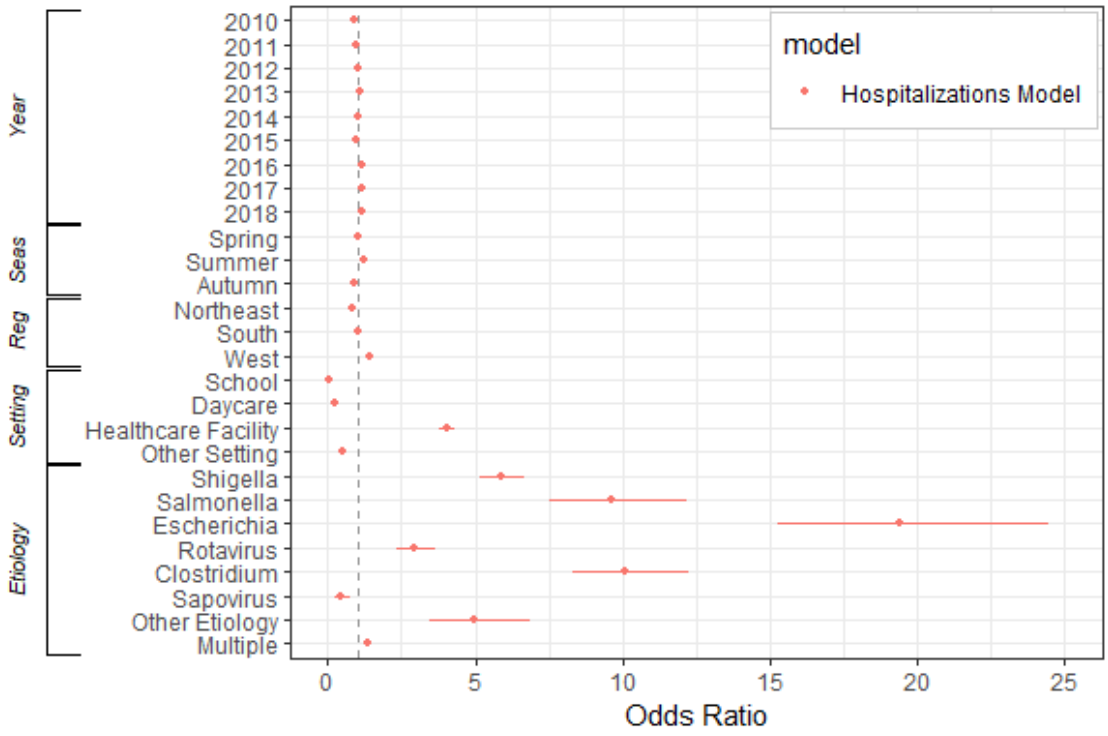

Figure 7 Multivariate Logistic Regression Model on Hospitalizations

4.2.2.3 Interpretation of Model Estimates

Due to the scarcity of the non-zero hospitalization counts, the confidence intervals are all noticeably larger in this logistic regression model for hospitalizations compared to the previous negative binomial model for illnesses.

Year multiplication factors for 2010-2015 contain 1.0 in the confidence intervals, meaning that the outbreaks in these years likely do not have different odds of hospitalization compared to outbreaks from 2009. Years 2016-2018 all had the same multiplication factor of 1.16, meaning that the odds of hospitalization from outbreaks in these years was, on average, 1.16 times that of outbreaks in 2009.

Spring (1.01) and Autumn (0.92) have multiplication factors rather close to 1.0, meaning the odds of hospitalization from outbreaks in these two seasons are not that different from outbreaks in winter. However, Summer has a multiplication factor of 1.23, meaning odds of hospitalization from outbreaks in summer were, on average, 1.23 times odds from outbreaks in winter.

Region is the first factor for which the levels start to show noticeably different probabilities for hospitalization. The Northeast region has a multiplication factor of 0.86, meaning that the odds of being hospitalized from an outbreak are lower in the Northeast compared to the Midwest holding all other factors constant. The West region has multiplication factor of 1.39, meaning that the odds of being hospitalized from an outbreak are, on average, 1.39 times higher in the West compared to the Midwest.

Settings also greatly influence probabilities of hospitalization. School (0.07) and Child Daycare (0.21) have very low multiplication factors, meaning outbreaks in these settings have much lower probabilities of having people be hospitalized than outbreaks in long term care facilities. Healthcare facilities had a relatively high multiplication factor at 4.05, meaning that the odds of hospitalization from outbreaks in healthcare facilities are, on average, 4 times greater than that of long term care facilities.

Each etiology appears to have very different multiplication factors for hospitalization. With Norovirus as the baseline, all of the other etiologies have odds ratios greater than 1.0 except for Sapovirus at 0.44. Most notably, *Salmonella* (9.64), *Clostridium* (10.10), and *Escherichia* (19.40) appear to have outbreaks for which the odds of being hospitalized are many times more likely than outbreaks of Norovirus.

Similar to the illness model, setting and etiology appear to be the factors with the most impact on probability of hospitalization.

4.2.3 Deaths

Because death was such a rare occurrence among the outbreaks in the dataset, the distribution of the number of deaths for the outbreaks was examined in Table 16.

Table 16 Frequency of Death Counts

Number of Deaths	Number of Outbreaks (Percentage)
0	19598 (85.52%)
1	549 (2.40%)
2	98 (0.43%)
3	31 (0.14%)
4	7 (0.03%)
5	3 (0.01%)
6	5 (0.02%)
7	1 (0.004%)
NA	2625 (11.45%)

Table 16 shows that 85.52% of the outbreaks had 0 deaths. There were only 8 possible death counts excluding missing values.

Due to this relative scarcity, regression models were not considered for the death outcome. Instead, descriptive statistics for setting and etiology (the two factors found to be the most

influential from the negative binomial model for illness count and the logistic regression model for hospitalization probability) were examined for outbreaks with non-zero death count values.

4.2.3.1 Setting

The settings for outbreaks with deaths were examined in Table 17.

Table 17 Outbreaks with Deaths by Setting

Setting	Number of Outbreaks with Deaths	Number of Outbreaks in Setting	Percentage of Fatal Outbreaks in Setting
Long Term Care Facility	624	13316	4.69%
School/College/University	2	2156	0.09%
Child Daycare	2	1492	0.13%
Healthcare Facility	20	832	2.40%
Other	8	1672	0.48%
Unknown	38	3449	1.10%
Total	694	22917	

Table 17 shows that almost all of the outbreaks with deaths were from the long term care facility setting. The percentage of having at least one death was very low at below 5% for each setting.

Pearson's chi-square test of independence was used to examine if there was a relationship between setting and whether or not an outbreak resulted in death(s). This test resulted in a p-value < 0.0001, meaning that there is a significant relationship between setting and death in this dataset.

4.2.3.2 Etiology

The etiologies for outbreaks with deaths were examined in Table 18.

Table 18 Outbreaks with Deaths by Etiology

Etiology	Number of Outbreaks with Deaths	Number of Outbreaks by Etiology	Percentage of Fatal Outbreaks by Etiology
Norovirus	543	15530	3.50%
Shigella	1	854	0.12%
Salmonella	5	219	2.28%
Clostridium	7	164	4.27%
Escherichia	3	138	2.17%
Rotavirus	6	104	5.77%
Sapovirus	1	78	1.28%
Other	1	182	0.55%
Multiple	32	611	5.24%
Unknown	95	5037	1.89%
Total	694	22917	

Table 18 shows that almost all of the outbreaks with deaths were from the Norovirus etiology. Despite this, Rotavirus outbreaks and Multiple etiology outbreaks had the highest percentages of fatal outbreaks.

Pearson's chi-square test of independence was used to examine if there was a relationship between etiology and whether or not an outbreak resulted in death(s). This test resulted in a p-

value of less than 0.0001, meaning that there is a significant relationship between etiology and death in this dataset.

4.2.4 Relationship Between Setting and Etiology

As different pathogens are known to thrive in different environments, the most common settings for each etiology were examined. Table 19 shows those settings and their percentages for each etiology.

Table 19 Most Common Settings for Outbreak Etiologies

Etiology	Most Common Setting (Percentage)	Second Most Common Setting (Percentage)
Norovirus	LTCF (73.41%)	School (10.74%)
Shigella	Daycare (59.11%)	School (22.17%)
Salmonella	Other Setting (40.79%)	Daycare (39.47%)
Clostridium	LTCF (66.67%)	Healthcare Facility (24.24%)
Escherichia	Daycare (63.36%)	Other Setting (25.19%)
Rotavirus	LTCF (73.96%)	Daycare (18.75%)
Sapovirus	LTCF (75.34%)	School (12.33%)
Other	Daycare (48.50%)	Other Setting (41.92%)
Mulitple	LTCF (59.38%)	Daycare (13.94%)

Table 19 shows that long term care facilities are usually the most common setting for outbreaks of different etiologies. However, *Shigella* outbreaks, *Escherichia* outbreaks, and Other

Etiology outbreaks are all most common in daycares. Schools and daycares are also the second most common setting for different etiologies.

Pearson's chi-square test of independence was used to examine if there was a relationship between setting and etiology. This test resulted in a p-value of less than 0.0001, meaning that there is a significant relationship between setting and etiology in this dataset.

4.3 Results from Multiple Imputation

4.3.1 Hot-Deck Imputation

Year, season, region, setting, etiology, and number of illnesses were used to impute missing values in setting and etiology. Chi-square tests for each of these factors confirmed that all of them were related to both setting and etiology. For each outbreak with missing setting and/or etiology values, similar complete outbreaks in other fully observed variables were found and used to create neighborhoods for imputations.

As year, season, region, setting, and etiology were categorical variables, similar outbreaks selected for the neighborhoods were required to have exactly matching values for the fully observed variables.

Unlike the other variables, illness count was a continuous variable and therefore needed be standardized before comparison. Raw illness counts were log-transformed, truncated by mean plus or minus three times its standard deviation, then divided by the range after truncation. The difference between these standardized values was used as the second metric to select outbreaks

with complete records to form the neighborhoods for outbreaks with missing data. The difference in the standardized scale was required to be less than or equal to 0.1.

Criteria were relaxed if there were 0 complete-case observations that matched the above requirements. They were modified such that one of year, season, region, setting, or etiology does not need to match (the rest still must match) while the difference between the two standardized illness values was increased to be less than or equal to 0.2 instead of 0.1.

4.3.2 Multiple Imputation

Using the Hot-Deck Imputation method described in the previous section, each outbreak with missing setting and/or missing etiology values was imputed 100 times by sampling with replacement from the neighborhood of similar observations, forming 100 different imputed datasets.

The same model used to fit the complete-case analysis for each outcome was used to run regressions on each of the 100 different datasets. Model coefficients were averaged over the 100 regressions. Variance was computed using the following formula:

$$T_D = \bar{W}_D + \frac{D+1}{D} B_D$$

Where T_D is the total variance, $\bar{W}_D$ is the average within-imputation variance, B_D is the between imputation variance, and D is 100. This variance was used to construct confidence intervals and p-values using a normal distribution for each coefficient.

These estimates were compared to the estimates from the complete-case analyses to examine the effect of the imputation on the data.

4.3.2.1 Illnesses

The negative binomial regression model with year, season, region, setting and etiology as covariates and counts of illness as the outcome used for the complete-case analysis was used for multiple imputation. Table 20 compares the covariate estimates for the multiple imputation analysis and the complete-case analysis.

Table 20 Negative Binomial Regressions for Multiple Imputation Analysis and Complete-Case Analysis

		Multiple Imputation		Complete-Case	
Covariate		Multiplication Factor [95% CI]	P-value	Multiplication Factor [95% CI]	P-value
Year					
	2009	1.00		1.00	
	2010	1.10 [1.04, 1.16]	0.0012	1.04 [0.96, 1.13]	0.3298
	2011	1.07 [1.01, 1.14]	0.0140	1.07 [0.99, 1.16]	0.0972
	2012	0.99 [0.94, 1.05]	0.7499	0.94 [0.87, 1.01]	0.0949
	2013	0.92 [0.87, 0.97]	0.0015	0.82 [0.77, 0.89]	< 0.0001
	2014	0.94 [0.89, 0.99]	0.0203	0.88 [0.82, 0.95]	0.0012
	2015	0.91 [0.86, 0.96]	0.0006	0.84 [0.78, 0.90]	< 0.0001
	2016	0.82 [0.78, 0.86]	< 0.0001	0.73 [0.68, 0.77]	< 0.0001
	2017	0.78 [0.74, 0.82]	< 0.0001	0.71 [0.67, 0.77]	< 0.0001
	2018	0.74 [0.70, 0.78]	< 0.0001	0.65 [0.61, 0.70]	< 0.0001
Season					
	Winter	1.00		1.00	
	Spring	0.87 [0.85, 0.89]	< 0.0001	0.88 [0.86, 0.91]	< 0.0001
	Summer	0.77 [0.73, 0.81]	< 0.0001	0.81 [0.76, 0.86]	< 0.0001
	Autumn	0.91 [0.88, 0.94]	< 0.0001	0.92 [0.88, 0.96]	0.0002
Region					
	Midwest	1.00	< 0.0001	1.00	

Northeast	1.09 [1.06, 1.12]	< 0.0001	1.11 [1.07, 1.15]	< 0.0001
South	1.09 [1.06, 1.12]	< 0.0001	1.06 [1.03, 1.09]	0.0005
West	1.09 [1.05, 1.13]	< 0.0001	1.07 [1.03, 1.12]	0.0009
Setting				
Long Term Care Facility	1.00		1.00	
School/College/University	1.84 [1.77, 1.91]	< 0.0001	1.84 [1.77, 1.92]	< 0.0001
Child Daycare	0.67 [0.64, 0.71]	< 0.0001	0.64 [0.60, 0.68]	< 0.0001
Healthcare Facility	0.68 [0.65, 0.72]	< 0.0001	0.70 [0.66, 0.75]	< 0.0001
Other	0.98 [0.94, 1.02]	0.3950	1.04 [0.99, 1.10]	0.0744
Etiology				
Norovirus	1.00		1.00	
Shigella	0.68 [0.64, 0.71]	< 0.0001	0.68 [0.63, 0.73]	< 0.0001
Salmonella	0.22 [0.19, 0.24]	< 0.0001	0.25 [0.21, 0.29]	< 0.0001
Clostridium	0.33 [0.29, 0.38]	< 0.0001	0.39 [0.32, 0.48]	< 0.0001
Escherichia	0.33 [0.28, 0.38]	< 0.0001	0.32 [0.28, 0.38]	< 0.0001
Rotavirus	0.68 [0.59, 0.79]	< 0.0001	0.67 [0.57, 0.80]	< 0.0001
Sapovirus	0.82 [0.70, 0.97]	0.0214	0.86 [0.72, 1.03]	0.0986
Other	0.27 [0.23, 0.30]	< 0.0001	0.25 [0.22, 0.29]	< 0.0001
Multiple	1.31 [1.23, 1.40]	< 0.0001	1.35 [1.26, 1.44]	< 0.0001

The coefficient estimates from Table 20 are displayed with a dot and whisker plot in Figure 8.

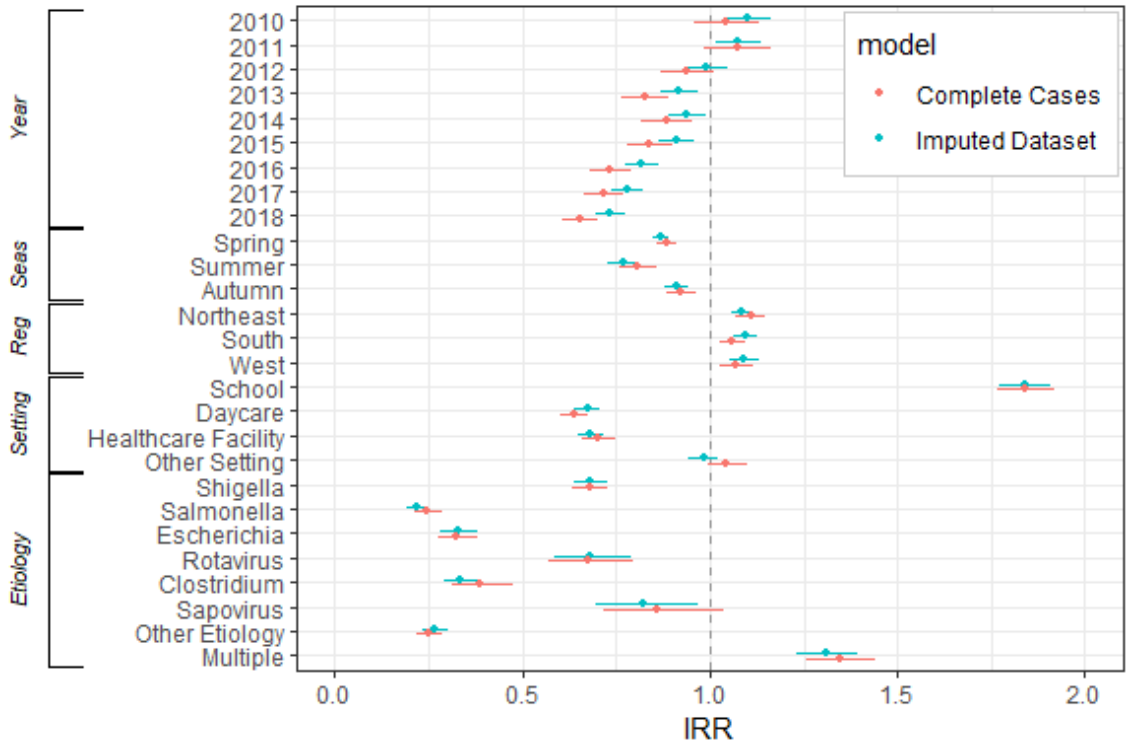

Figure 8 Negative Binomial Regressions for Multiple Imputation Analysis and Complete-Case Analysis

Table 20 and Figure 8 show that the coefficient estimates from the multiple imputation analysis are very close to those of the complete case analysis. The most notable difference is that the estimates from the multiple imputation analysis have narrower confidence intervals, due to a reduction in variance by the inclusion of information from incomplete observations in the regression analysis.

4.3.2.2 Hospitalizations

The logistic regression model with year, season, region, setting and etiology as covariates and hospitalizations vs not hospitalizations as the outcomes used for the complete-case analysis

was used for multiple imputation. Table 21 compares the covariate estimates for the multiple imputation analysis and the complete-case analysis.

Table 21 Logistic Regressions for Multiple Imputation Analysis and Complete-Case Analysis

		Mulitple Imputation		Complete-Case	
Covariate		Multiplication Factor [95% CI]	P-value	Mulitplication Factor [95% CI]	P-value
Year					
	2009	1.00		1.00	
	2010	0.89 [0.80, 0.99]	0.0357	0.90 [0.79, 1.02]	0.0949
	2011	0.90 [0.81, 1.00]	0.0494	0.95 [0.84, 1.08]	0.4394
	2012	0.97 [0.88, 1.08]	0.6168	1.05 [0.94, 1.18]	0.3863
	2013	0.98 [0.89, 1.09]	0.7287	1.07 [0.95, 1.20]	0.2489
	2014	1.00 [0.90, 1.10]	0.9659	1.00 [0.89, 1.12]	0.9898
	2015	1.01 [0.91, 1.11]	0.8911	0.96 [0.86, 1.08]	0.5220
	2016	1.17 [1.06, 1.30]	0.0017	1.16 [1.04, 1.30]	0.0120
	2017	1.12 [1.02, 1.25]	0.0329	1.16 [1.04, 1.30]	0.0106
	2018	1.11 [1.00, 1.23]	0.0518	1.16 [1.03, 1.30]	0.0127
Season					
	Winter	1.00		1.00	
	Spring	0.98 [0.94, 1.03]	0.4969	1.01 [0.96, 1.07]	0.5901
	Summer	1.12 [1.01, 1.25]	0.0329	1.23 [1.10, 1.38]	0.0003
	Autumn	0.89 [0.83, 0.95]	0.0008	0.92 [0.86, 1.00]	0.0415
Region					
	Midwest	1.00		1.00	
	Northeast	0.89 [0.84, 0.94]	< 0.0001	0.86 [0.81, 0.91]	< 0.0001
	South	1.07 [1.02, 1.12]	0.0086	1.02 [0.97, 1.08]	0.4311
	West	1.47 [1.38, 1.56]	< 0.0001	1.39 [1.30, 1.49]	< 0.0001

Setting				
Long Term Care Facility	1.00		1.00	
School/College/University	0.10 [0.09, 0.11]	< 0.0001	0.07 [0.06, 0.08]	< 0.0001
Child Daycare	0.25 [0.22, 0.29]	< 0.0001	0.21 [0.18, 0.24]	< 0.0001
Healthcare Facility	3.47 [3.25, 3.70]	< 0.0001	4.05 [3.78, 4.34]	< 0.0001
Other	0.61 [0.56, 0.67]	< 0.0001	0.51 [0.46, 0.56]	< 0.0001
Etiology				
Norovirus	1.00		1.00	
Shigella	4.49 [3.92, 5.14]	< 0.0001	5.86 [5.14, 6.68]	< 0.0001
Salmonella	7.72 [6.06, 9.85]	< 0.0001	9.64 [7.54, 12.18]	< 0.0001
Clostridium	8.49 [6.70, 10.76]	< 0.0001	10.10 [8.31, 12.21]	< 0.0001
Escherichia	14.90 [11.68, 19.01]	< 0.0001	19.40 [15.23, 24.48]	< 0.0001
Rotavirus	2.46 [1.94, 3.13]	< 0.0001	2.96 [2.35, 3.67]	< 0.0001
Sapovirus	0.55 [0.31, 0.97]	0.0382	0.44 [0.23, 0.74]	0.0044
Other	4.32 [3.13, 5.97]	< 0.0001	4.96 [3.48, 6.88]	< 0.0001
Multiple	1.31 [1.19, 1.45]	< 0.0001	1.36 [1.23, 1.49]	< 0.0001

The coefficient estimates from Table 21 are displayed with a dot and whisker plot in Figure 9.

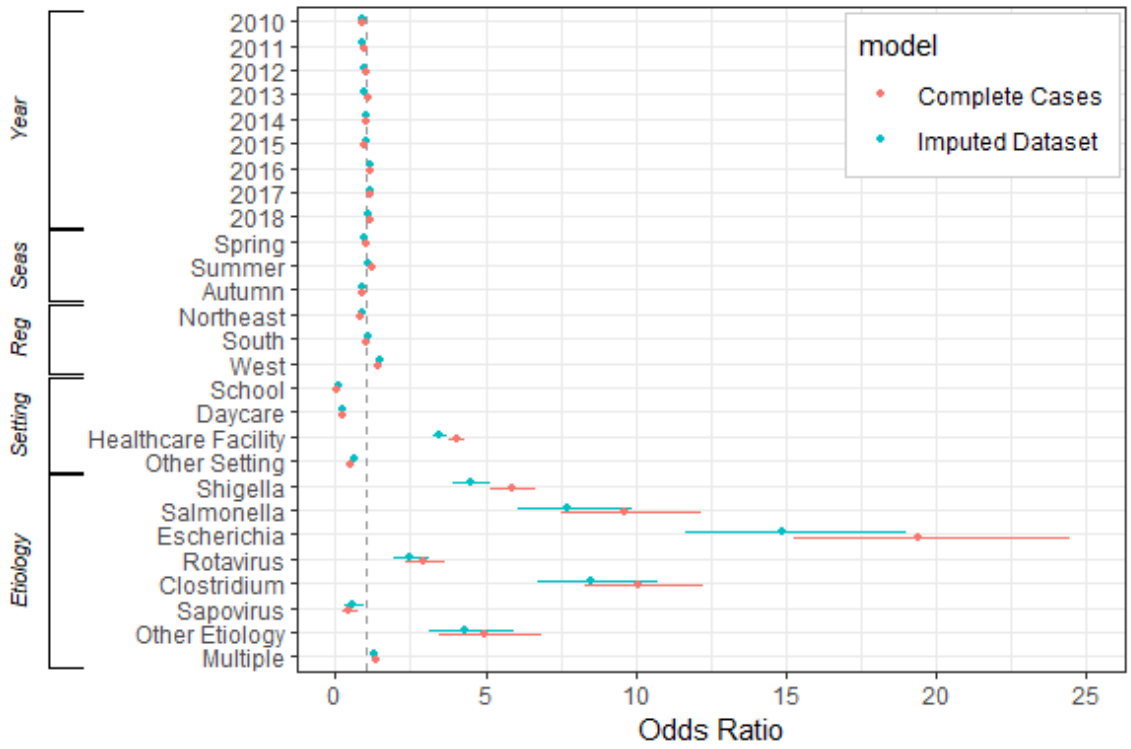

Figure 9 Logistic Regressions for Multiple Imputation Analysis and Complete-Case Analysis

Table 21 and Figure 9 show that the coefficient estimates for year, season, region, and setting from the multiple imputation analysis are very close to those of the complete case analysis. The most noticeable differences come from the coefficient estimates for etiology. This makes sense as most of the missing values from the dataset were from the etiology variable. Once again, the estimates from the multiple imputation analysis have narrower confidence intervals, due to a reduction in variance by the inclusion of information from incomplete observations in the regression analysis.

5.0 Discussion

In this thesis, we examined the relationship between year, season, region, setting, and etiology and the outcomes of illness, hospitalization, and death. A negative binomial regression model was used to study how the number of illnesses in an enteric disease outbreak varied across those factors. A logistic regression model was used to examine how the chance of hospitalization among affected individuals in an enteric disease outbreak depended on those factors. Because death only occurred in very few outbreaks, the Pearson's chi-square test was used to study the association between each individual factor and presence of death among the infected in an enteric disease outbreak.

Descriptive statistics of factors of interest in this dataset matched those of prior studies. Winter was the most common season for outbreaks. Norovirus outbreaks were the most common with regards to etiology.

Results from the regression analyses on the number of illnesses indicate that the average numbers of illnesses among enteric disease outbreaks in schools and long term care facilities are much higher than outbreaks in other settings. On the contrary, outbreaks under child daycare and healthcare facility had fewer numbers of illnesses (Table 9). These findings agree with previous studies that found that Norovirus outbreaks in particular are uniquely suited to areas of close living quarters, shared dining facilities, and difficult environmental maintenance (Robilotti, 2015). As a majority of the outbreaks in this dataset involve Norovirus, these setting trends are expected. Schools and long term care facilities tend to fit these three criteria listed better than the other settings examined in the regression analysis, so it makes sense for these settings to have greater numbers of illnesses. It was also noted that enteric disease outbreaks caused by Norovirus or

multiple pathogens (can be interpreted as mostly Norovirus with at least one other etiology) were the ones that led to much higher numbers of infected people. This agrees with research that has found that Norovirus is very contagious as it can stay on surfaces even after disinfection and cause infection with as few as ten viral particles (Capece, 2020). Outbreaks caused by *Salmonella*, *Clostridium*, and *Escherichia* tended to have fewer numbers of illnesses.

Results from the regression analyses on the probability of hospitalization indicate that outbreaks in settings of school and daycare have a much lower probability of resulting in hospitalizations compared to outbreaks in long term care facilities. This result supports previous findings that found that implementation of Rotavirus vaccines in 2008 appear to have significantly reduced the number of child hospitalizations from acute gastroenteritis in the United States (Leshem, 2018). This makes sense as children are historically the demographic most at risk for Rotavirus in particular. Healthcare facilities have a much higher probability of hospitalization, but this might be a technicality because hospitals are included in this categorization. With regards to etiology, outbreaks with Norovirus and Sapovirus had low probabilities of resulting in hospitalization relative to the other etiologies. This result supports prior research that found that symptoms of Sapovirus infection are usually mild and patients tend to recover within a couple of days, similarly Norovirus infections (Oka, 2015). *Salmonella*, *Escherichia*, and *Clostridium* in particular had very high probabilities of hospitalization relative to other etiologies. This is notable as outbreaks with these same three etiologies were found to often result in low numbers of illnesses, implying that outbreaks with these three etiologies often result in more severe outcomes relative to the other etiologies in this dataset. Most previous publications on person-to-person transmitted enteric disease outbreaks focus on single pathogens individually, making it difficult to say whether this finding is supported by the literature. However, these pathogens have all been

listed individually on multiple occasions as notable causes of concern for severe outcomes in analyses on foodborne outbreaks, meaning this finding makes sense in context.

Descriptive statistics for fatal outbreaks showed that few outbreaks with deaths occurred. Those that did occur only all had very low death counts. Almost all of these outbreaks occurred in long term care facilities. With regards to etiology, a large majority of the outbreaks with deaths were related to Norovirus. This is expected as the majority of all outbreaks in the dataset were in these two categories as well. Rotavirus outbreaks and multiple etiology outbreaks in particular had relatively high percentages of resulting in at least one death relative to the other etiologies. Chi-square tests of independence found both setting and etiology to be significantly associated with mortality.

The examination of the relationship between setting and etiology of outbreaks found that common settings differed based on the pathogen causing the outbreak. The result of the chi-square test indicated a significant relationship between these two factors. Long term care facilities and daycares in particular were often the settings with the highest frequencies across etiologies, implying that the most vulnerable populations were either very young or very old individuals depending on the pathogen causing the outbreak. Norovirus outbreaks appear to be especially prevalent in older adults in long term care facilities, a finding supported by previous research (Chen, 2017). A majority of *Shigella* outbreaks and *Escherichia* outbreaks occurred in daycares, which is consistent with the CDC website's statements (https://www.cdc.gov/shigella/infection-sources.html) that young children are more likely to get infected and develop complications by these two pathogens (https://www.cdc.gov/healthypets/diseases/ecoli.html).

Multiple imputation succeeded in reducing the variance of the coefficient estimates in both models by imputing missing setting and missing etiology values. The estimates from the multiple

imputation analysis were rather similar with noticeably narrower confidence intervals, indicating improved precision from the multiple imputation. With the results from hospitalizations model in particular, some categories had very few outbreaks with nonzero hospitalization counts. Therefore, the multiple imputation estimates were shown to differ more from the complete case estimates compared to the illnesses model.

The dataset extracted from NORS contained many variables, but only a few were usable for the regression models. This was because many variables were details on other variables in the dataset. More distinct variables would have been helpful in modeling the outcomes. Multiple imputation was able to address the outbreaks with missing values in setting and etiology, but was not able to account for the majority of outbreaks with zero hospitalizations and outbreaks with zero deaths. This scarcity of outbreaks with more severe outcomes made examining hospitalization and death from the dataset particularly difficult.

In the hot-deck imputation procedure, data on year, region, season, setting, etiology, and illness count were incorporated to impute missing values in etiology and settings of many outbreaks. It is noted that the imputation was based on the condensed levels of season, region, setting, and etiology, meaning that some information from the original data was lost. The imputation procedure could potentially be improved by using the more granular data from the variables from the original dataset before reclassification. For example, we could match potential donors with the outbreak with missing values by month instead of season, by state instead of region, and by the specific setting instead of the combined setting. This would provide more specificity for the hot-deck imputation procedure, leading imputations to more closely resemble similar complete-case donor observations. The only complication would be vastly increased number of strata as determined by the matching variables in more granular levels, potentially

making it difficult to find any donors within the same stratum as the outbreak to be imputed. In order to identify a few donors that are reasonably close to the outbreak under consideration, one could develop a strategy to collapse on some strata and identify a neighborhood with donors that are exactly matched on some variables and similar in other variables to the target outbreak. Improvements on imputation could lead to even further reductions on variation, which would result in narrower confidence intervals for model coefficient estimates.

The models used to examine illness counts and hospitalization probabilities were both main effects models. Interactions between covariates were considered, but ultimately not included in the modeling process because including the resulting models suffered from sparsity. The negative binomial regression model for illness fit the data relatively well but the logistic regression model for hospitalization was not adequate. Further research into improving these models could involve including more covariates, reorganizing covariates onto a continuous scale and/or refactoring covariates into fewer levels, so that interactions between covariates can be examined without sparsity or the risk of overfitting.

With Norovirus being such a common cause of acute gastroenteritis outbreaks in the United States, public health efforts should consider educating the public on basic cautionary measures and infection symptoms. As Norovirus spreads easily and quickly in different ways, the CDC recommends the following measures for prevention: practicing proper hand hygiene, handling and preparing food safely, not preparing food or caring for others when sick, disinfecting surfaces, and washing laundry thoroughly (https://www.cdc.gov/norovirus/about/prevention.html). Initiatives like these should be aimed towards younger people in school as acute gastroenteritis outbreaks tend to have higher illness counts on average in these settings. Efforts to mitigate hospitalizations from these outbreaks should be targeted towards long term care facilities and *Salmonella*,

Escherichia, and *Clostridium* in particular. The common thread between prevention advice for these three different pathogens is to make sure to wash hands after using the bathroom. Deaths are very rare from acute gastroenteritis outbreaks, but special attention should be given to Rotavirus outbreaks and multiple etiology outbreaks to prevent deaths.

Bibliography

About NORS | CDC. 1 Apr. 2021, https://www.cdc.gov/nors/about.html.

Agresti, Alan. Categorical Data Analysis, 3rd Edition. Hoboken, New Jersey, John Wiley & Sons, 2013.

Capece, Gregory, and Elizabeth Gignac. "Norovirus." StatPearls, StatPearls Publishing, 2021. PubMed, http://www.ncbi.nlm.nih.gov/books/NBK513265/

CDC. "Preventing Norovirus." Centers for Disease Control and Prevention, 5 Mar. 2021, https://www.cdc.gov/norovirus/about/prevention.html.

Chen, Yingxi, et al. "Norovirus Disease in Older Adults Living in Long-Term Care Facilities: Strategies for Management." Current Geriatrics Reports, vol. 6, no. 1, 2017, pp. 26–33. PubMed Central, doi:10.1007/s13670-017-0195-z.

Christopher, Prince RH, et al. "Antibiotic Therapy for Shigella Dysentery." The Cochrane Database of Systematic Reviews, vol. 2010, no. 8, Aug. 2010. PubMed Central, doi:10.1002/14651858.CD006784.pub4.

Cook, R. D. and Weisberg, S. Residuals and Influence in Regression. London, Chapman and Hall. 1982

Elliott, Elizabeth Jane. "Acute Gastroenteritis in Children." BMJ : British Medical Journal, vol. 334, no. 7583, Jan. 2007, pp. 35–40. PubMed Central, doi:10.1136/bmj.39036.406169.80.

E. Coli Infection | Healthy Pets, Healthy People | CDC. 8 July 2019, https://www.cdc.gov/healthypets/diseases/ecoli.html.

Fares, Auda. "Factors Influencing the Seasonal Patterns of Infectious Diseases." International Journal of Preventive Medicine, vol. 4, no. 2, Feb. 2013, pp. 128–32.

Forms & Guidance | National Outbreak Reporting System (NORS) | CDC. 11 Mar. 2021, https://www.cdc.gov/nors/forms.html

Graves, Nancy S. "Acute Gastroenteritis." Primary Care, vol. 40, no. 3, Sept. 2013, pp. 727–41. PubMed, doi:10.1016/j.pop.2013.05.006.

Hall, Aron J., et al. "Acute Gastroenteritis Surveillance through the National Outbreak Reporting System, United States." Emerging Infectious Diseases, vol. 19, no. 8, Aug. 2013, pp. 1305–09. PubMed Central, doi:10.3201/eid1908.130482.

Leshem, Eyal, et al. “National Estimates of Reductions in Acute Gastroenteritis-Related Hospitalizations and Associated Costs in US Children After Implementation of Rotavirus Vaccines.” Journal of the Pediatric Infectious Diseases Society, vol. 7, no. 3, Aug. 2018, pp. 257–60. PubMed, doi:10.1093/jpids/pix057.

Little, Roderick J. A. and Rubin, Donald B. Statistical Analysis with Missing Data, Second Edition. John Wiley & Sons, 2002.

Lopman, Ben, et al. “Host, Weather and Virological Factors Drive Norovirus Epidemiology: Time-Series Analysis of Laboratory Surveillance Data in England and Wales.” PLoS ONE, vol. 4, no. 8, Aug. 2009. PubMed Central, doi:10.1371/journal.pone.0006671.

Morillo, Simone Guadagnucci, and Maria do Carmo Sampaio Tavares Timenetsky. “Norovirus: An Overview.” Revista Da Associacao Medica Brasileira (1992), vol. 57, no. 4, Aug. 2011, pp. 453–58. PubMed, doi:10.1590/s0104-42302011000400023.

Papadopoulos, Theofilos, et al. “The Health and Economic Impact of Acute Gastroenteritis in Belgium, 2010-2014.” Epidemiology and Infection, vol. 147, Jan. 2019, p. e146. PubMed, doi:10.1017/S095026881900044X.

Robilotti, Elizabeth, et al. “Norovirus.” Clinical Microbiology Reviews, vol. 28, no. 1, Jan. 2015, pp. 134–64. PubMed, doi:10.1128/CMR.00075-14.

Rubin, Donald B., Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, 1987.

Sources of Infection & Risk Factors | Shigella – Shigellosis | CDC. 17 Jan. 2019, https://www.cdc.gov/shigella/infection-sources.html.

Wielgos, Katarzyna, et al. “[Management of acute gastroenteritis in children].” Polski Merkuriusz Lekarski: Organ Polskiego Towarzystwa Lekarskiego, vol. 47, no. 278, Aug. 2019, pp. 76–79.

Wikswo, Mary E., et al. “Outbreaks of Acute Gastroenteritis Transmitted by Person-to-Person Contact, Environmental Contamination, and Unknown Modes of Transmission--United States, 2009-2013.” Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002), vol. 64, no. 12, Dec. 2015, pp. 1–16. PubMed, doi:10.15585/mmwr.mm6412a1.

Williams, D. A. “Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions.” Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 36, no. 2, 1987, pp. 181–191. JSTOR, www.jstor.org/stable/2347550. Accessed 18 June 2021.