2017

Analysis of Food Exposures in Foodborne Disease Outbreaks

Alecia T. Alianell
University of South Carolina

Analysis of Food Exposures in Foodborne Disease Outbreaks

by

Alecia T. Alianell

Bachelor of Science
University of Southern California, 2006

Master of Public Health
Brown University, 2008

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Epidemiology

The Norman J. Arnold School of Public Health

University of South Carolina

2017

Accepted by:

Anwar Merchant, Major Professor

Eric Brenner, Committee Member

Alexander McLain, Committee Member

Dana Giurgiutiu, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

DEDICATION

This is dedicated to my parents, Dr. Gary Alianell and Michele Alianell, my sister, Elizabeth Alianell, and of course Elmer, for their support throughout this process.

ACKNOWLEDGEMENTS

I would like to acknowledge my dissertation committee chair, Dr. Anwar Merchant, for his guidance and support as well as passing along his experience in epidemiology, nutrition, and foodborne illness research. I would also like to acknowledge Dr. Alexander McLain for all of his knowledge and expertise in the field of biostatistics, Dr. Eric Brenner for imparting his experience in epidemiology and infectious diseases, and Dr. Dana Giurgiutiu for conveying her experience in acute disease epidemiology. I would also like to acknowledge David Young for all of his help with retrieving data from DHEC.

ABSTRACT

Foodborne illness remains a serious public health problem in the United States in general as well as South Carolina in particular. Obtaining good food ingestion histories as well as possible risky environmental exposures is one of the earliest, most important tasks to complete in any foodborne outbreak investigations. Because time is of the essence in investigations, we have evaluated a rarely used biostatistical method, *Random Forests,* to data obtained from DHEC. *Random Forests* has the potential to facilitate more rapid identification of foods or environmental exposures that may be associated with outbreaks. We also examined previous cases of salmonellosis using two different definitions (state and FDA) of what constitutes a foodborne outbreak using logistic regression with a Poisson distribution. Dietary patterns were similarly evaluated, as they are associated with mortality from all causes. We aimed to characterize the nutrition and dietary intake of South Carolina residents and see what foods eaten may be associated with foodborne outbreaks. In summary, we have used *Random Forests* to analyze data that are routinely collected during foodborne outbreak investigations. This new application *of Random Forests* can make identification of foods responsible outbreaks more efficient. This information will address the challenges of a rural southern state with a high obesity rate by using a representative sample that contains geographic and socio-demographic diversity and using said information to help affect change in the programs available. The results of this study can potentially improve foodborne disease outbreak investigations in South Carolina.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DHEC............................South Carolina Department of Health and Environmental Control

IPSPR......University of South Carolina's Institute for Public Service and Policy Research

STEC...Shiga toxin-producing Escherichia coli

FDA..Food and Drug Administration

CDC ....Centers for Disease Control and Prevention

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

With foodborne outbreaks constituting a major, ongoing public health burden in South Carolina and the United States, prompt and effective detection of the source through epidemiologic investigations are necessary to remove contaminated food from the market. These investigations help prevent further illnesses, and focus prevention strategies on critical contamination points along the "farm-to-fork" continuum. The plan, in conjunction with South Carolina's Department of Environmental Health and Control (DHEC) is to enhance detection, investigation, and control of foodborne disease.

Currently, it is mandated by the state of South Carolina that cases of infection with *Salmonella* need to be reported to DHEC within three business days. While many more pathogens cause gastrointestinal illness, the main focus of the present research will be *Salmonella* as it is one of the top causes of foodborne illness in the US. DHEC staff monitors daily laboratory and provider disease reports to identify positive reports, followed by immediate interview of those with positive laboratory results. For the interview, a standard core questionnaire assessing food history was developed and implemented by DHEC for initial screening and hypothesis generation. Interviews are administered without waiting for the serotype results. Conducting real-time review of subtyping results in conjunction with the interviews can make it possible to identify cluster-associated cases.

But, while the interviews are being conducted, organisms are sent to the DHEC Bureau of Laboratories (BoL) to be characterized. These results were used to separate the interviews into cases possibly associated with a common source of exposure versus sporadically occurring cases. The incidence, rates, and risk of certain pathogens in the state as a whole as well as by geographical area can be determined from the questionnaires through the use of a cohort study. Currently, foods associated with an outbreak are characterized using excel to manually look for possible associations.

Eighteen counties (Pickens, Greenville, Spartanburg, Cherokee, Union, Richland, Lexington, Newberry, Fairfield, Chester, Lancaster, York, Horry, Williamsburg, Georgetown, Berkeley, Dorchester, and Charleston) were selected for this sentinel project. This was based upon: 1) disease burden (66% of all cases of *Salmonella*, Shiga toxin-producing Escherichia coli (STEC) and Listeria), 2) geographic diversity (Upstate, Midlands and Coastal areas), 3) population density (these counties represent 65% of the total state population), and, 4) presence of three metropolitan statistical areas (MSAs) listed in the "Top 100" nationally within the selected counties.

To further the investigation into the food preferences of South Carolinians, the DHEC Food Exposures Survey was conducted in conjunction with the University of South Carolina's Institute for Public Service and Policy Research (IPSPR). This is a telephone (both landline and cell phone) survey with a representative sample of the state's population two years of age or older that will help determine the level of exposure that South Carolinians have to various types of food. The questionnaire used in this survey was designed based on the previously mentioned hypothesis-generating questionnaire that is used by DHEC when it encounters a confirmed case of *Salmonella*. Each of the questionnaires was administered to a representative sample of the target population.

While these data contribute to disease surveillance, we plan to use them to characterize the dietary patterns of South Carolinians and test a new method to identify foods associated with outbreaks. This information will address the challenges of a rural southern state with high disease burden in locations with the highest number of cases, something that has not previously been shown.

1.2 SPECIFIC AIMS

The **Specific Aims** of this study are as followed:

Aim 1: To compare food(s) that may have caused a foodborne outbreak identified by the random forests method using data from standardized hypothesis-generating questionnaires conducted by the South Carolina Department of Health and Environmental Control.

Research Question 1.1: Does the random forests method applied to a foodborne outbreak study lead to identification of more defined food clusters?

Research Question 1.2: How do the groups identified by random forests compare to traditional methods (i.e. a case-only study)?

Aim 2: To evaluate food eaten by South Carolina residents who may have been involved in a foodborne outbreak using data from standardized hypothesis-generating questionnaires conducted by the South Carolina Department of Health and Environmental Control.

Research Question 2.1: Are there commonalities in dietary intake patterns among individuals who live in urban settings and those who live in rural settings and whether or not the individual has been involved in a foodborne outbreak?

Research Question 2.2: Is there a relation between food clusters identified by random forests and the dietary patterns of South Carolinians?

Aim 3: To characterize the nutrition and dietary intake of South Carolina residents using the South Carolina Department of Health and Environmental Control Food Exposures Survey.

Research Question 3.1: What are the demographic, social, and geographic determinants of dietary patterns in the South Carolina population not investigated for foodborne illness outbreaks?

Research Question 3.2: Is there an association between the dietary patterns and frequency of eating pre-packaged foods or fresh foods?

1.3 SIGNIFICANCE

While foodborne pathogens and the study of outbreaks themselves are not innovative, the ways they are being investigated continue to evolve. Random Forests is a relatively new data driven machine-learning tool to identify predictive patterns in big data that is used in many diverse fields. Its application to foodborne disease outbreak investigations can potentially help to identify foods causing illness quickly, but to the best of our knowledge, this has not yet been done. We have applied Random Forests to identify foods associated with *Salmonella* outbreaks in South Carolina using outbreak investigation data collected by DHEC (Aim 1). The results of this study can potentially improve foodborne disease outbreak investigations in South Carolina and beyond. Also, no other studies have looked at *Salmonella* in the context of regions in South Carolina.

In addition, we studied dietary patterns derived from questionnaires that DHEC administered to different groups of people in South Carolina. The first dietary pattern was derived from a food exposure questionnaire that had been administered to individuals as part of outbreak investigations. Food groups associated with *Salmonella* outbreaks were identified, and then described by demographic and other characteristics (Aim 2). The second food group pattern was derived from a questionnaire that was similar to the one DHEC had used for the outbreak investigation, but was administered to a representative sample of individuals living in South Carolina. We then described the demographic, social and geographic predictors of consumption of these food groups by South Carolina residents (Aim 3). This information characterizing food exposure data collected by DHEC as part of its surveillance, will help in interpreting data collected in outbreak investigations. To the best of our knowledge, no other study has attempted to collect this information and make this comparison. In summary, we have used Random Forests to analyze data that are routinely collected during foodborne outbreak investigations. This new application of Random Forests can make identification of foods responsible outbreaks more efficient.

CHAPTER 2

LITERATURE REVIEW

2.1 WHAT IS FOODBORNE ILLNESS?

Foodborne illness is a serious public health threat. The Centers for Disease Control and Prevention (CDC) estimates that 76 million foodborne illnesses, including 325,000 hospitalizations and 5,000 deaths, occur in the United States each year.[1,2] This roughly equates to 1 in 6 Americans getting sick every year.[2,3] Anyone can get a foodborne illness. However, some people are more likely to develop foodborne illnesses than others, including infants and children, pregnant women, older adults, and people with weak immune systems.[4]

Many different disease-causing microbes (or pathogens) can contaminate foods or beverages, leading to many different foodborne infections.[3] Most diseases are infections caused by a variety of bacteria, viruses, and parasites that can be foodborne. Since so many microbes can cause foodborne illness, there can be many different symptoms, but the most common are nausea, vomiting, abdominal cramps, and diarrhea.[3,5] Most healthy individuals will recover without treatment in about 4 to 7 days without the need for hospitalization.[6] Microbes can also spread in more than one way, so it can be tough to tell if a disease is foodborne, which is a matter of public health importance due to the fact that knowing how a disease is spread is crucial in the process of stopping it from spreading further.

2.1a *SALMONELLA*

One of the most common causes of foodborne illness is *Salmonella*. *Salmonella* is a Gram-negative facultative intracellular pathogen that causes a spectrum of clinical diseases depending on the serotype of the infecting bacteria and the susceptibility of the host.[7,8] Infections are classified in three categories, (1) gastroenteritis, (2) systemic infection of an otherwise healthy host, or typhoid, and (3) infection of an immunocompromised host. In terms of this review, focus will be on the first of three aforementioned types, gastroenteritis, or non-typhoidal *Salmonella*.

Large foodborne outbreaks, including those caused by *Salmonella*, associated with the ingestion of contaminated foods, like tomatoes, produce, and peanut butter are becoming a more and more common occurrence.[7,9-12] Approximately 50% of all foodborne infections are caused by bacteria, and of those, about 30-50% can be attributed to *Salmonella* and its variates.[1,13] What is important to note is that these intestinal infections can be initiated by any of the approximately 2,000 different serotypes of *Salmonella* that infect both human and animal reservoirs.[7]

The most common human isolates of non-typhoidal *Salmonella* (NTS) are *Salmonella* Enterica serotypes Typhimurium (*S.* Typhimurium) and Enteritidis (*S.* Enteritidis) as well as *S.* Newport *and S.* Heidelberg.[14] Annually, there are around 40,000 cases of NTS reported each year, which underestimates the actual problem due to the ill person not visiting a physician or the fact that no specimen is obtained for laboratory testing.[15,16] It is estimated that of these cases, 582 deaths occur each year, making this pathogen the leading cause of foodborne infections with lethal outcomes in the United

States.[14] The incidence of foodborne human infections caused by *S.* Enteritidis and by multi-drug-resistant strains of *S.* Typhimurium increased substantially during the second half of the 20th century moving into the first part of the 21st century in the US with similar trends being reported from Europe.[17-21]

2.1b NTS GLOBALLY

NTS is the single most common cause of death from diarrheal disease associated with viruses, parasites, or bacteria. In addition, it is the leading cause of foodborne disease outbreaks in the United States.[15,13] This produces between $500 million to $2.3 billion in annual costs for medical care and lost productivity.[22] In the US and Europe, the press does a good job in getting the word out about outbreaks, but what is less known publicly is the impact of NTS infections in developing countries. Diarrheal diseases result in approximately 2-3 million deaths among children annually in developing countries, of which a significant portion is caused by NTS.[14] NTS are currently the most common blood isolates from children and the second most common cause of neonatal meningitis in sub-Saharan Africa, resulting in mortality rates exceeding 20%.[23,24] In adults, NTS is associated with AIDS as a top risk factor due to the HIV epidemic. Annually, about 10% of HIV positive African adults develop NTS infections, resulting in mortality rates about 20%.[14]

2.2 MONITORING SYSTEMS

2.2a NATIONAL NOTIFIABLE DISEASES SURVEILLANCE SYSTEM (NNDSS)

The CDC National Notifiable Diseases Surveillance System (NNDSS) is a nationwide collaboration that enables all levels of public health (local, state, territorial, federal, and international) to share health information to monitor, control, and prevent the occurrence and spread of state-reportable and nationally notifiable infectious and some noninfectious diseases and conditions.[25] NNDSS is a multifaceted program that includes the system for collection, analysis, and sharing of health data. It also allows for the sharing of policies, laws, electronic messaging standards, people, partners, information systems, processes, and resources at the local, state, and national levels.[25] State health departments voluntarily submit notifiable disease information electronically and through the NNDSS; that information is collected, analyzed, interpreted, managed, and shared according to standards set by NNDSS. This information is then disseminated nationally through the Morbidity and Mortality Weekly Report (MMWR).

2.2b FOODCORE

While little is known publicly about the impact of NTS globally, the US does a good job of monitoring all outbreaks with a system called FoodCORE (Foodborne Diseases Centers for Outbreak Response Enhancement), which is a supplemental surveillance system to the NNDSS. In 2009, the CDC funded a pilot program to improve response to foodborne disease outbreaks in conjunction with U.S. Department of Agriculture's Food Safety and Inspection Service and the Association of Public Health Laboratories. Since it was successful, it was expanded to seven states, of which South Carolina is one, and renamed FoodCORE. This covers about 14% of the US population.[26] Each center works together to develop better methods to detect, investigate, respond to, and control multistate outbreaks of foodborne diseases. Efforts are focused on outbreaks caused by bacteria, including *Salmonella*. FoodCORE focuses on four key areas: enhancement of public health laboratory surveillance, epidemiologic interviews and investigations, environmental health assessments, and best practices and replicable models for detection, investigation, response, and control.[26]

2.2c PULSENET

Molecular subtyping of bacterial isolates has been successfully applied to help in epidemiologic investigations of foodborne disease outbreaks since plasmid fingerprinting was used close to 30 years ago.[27,28] Since then, several more methods for identifying restriction fragment length polymorphisms on chromosomal DNA have been developed, and molecular subtyping has become an integral part of epidemiologic investigations of infectious diseases, including foodborne outbreaks.[29-36] Given the number of techniques and protocols that can be used for subtyping, even for the same type of bacteria, the results cannot be compared across laboratories, leading to diminished power and thus diminished identification of outbreaks.

In 1993, there was an investigation of an *Escherichia coli* O157:H7 outbreak caused by contaminated hamburgers served in a fast-food restaurant chain in the western United States.[28] Barrett et al. applied pulsed-field gel electrophoresis (PFGE) to characterize clinical and food isolates of *E. coli* O157:H7 and demonstrated its utility in outbreak investigations.[37] This led to a high demand in requests, which in turn led to the realization that decentralization of subtyping activities to public health laboratories would enable more timely subtyping of food isolates. This would have information more readily available to epidemiologists while investigating outbreaks.

Thus, in 1995, the CDC, with assistance from the Association of Public Health Laboratories (APHL) created a national molecular subtyping network for foodborne bacterial disease surveillance, later to be known as PulseNet.[28,38] Over the years, PulseNet's laboratory evaluation of isolates from clusters or outbreaks identified through epidemiologic surveillance has demonstrated its value in early recognition of outbreaks and rapid identification of their microbial sources.

2.3 EXAMPLES OF *SALMONELLA* OUTBREAKS IN THE US

2.3a *SALMONELLA* SAINTPAUL

There have been many national outbreaks that were cause for concern in the United States in the past decade. Raw produce is an increasingly recognized vehicle for transmission of pathogens.[39-41] In May 2008, the New Mexico Department of Health notified the CDC of 19 cases of *salmonella* infection, of which 7 had completed serotyping. All 7 came back as *Salmonella* Enterica serotype Saintpaul. Four isolates tested had indistinguishable patterns on PFGE. Later in the month, PulseNet staff identified three additional isolates in Colorado and Texas, which was surprising considering only 40 human Saintpaul isolates were submitted to PulseNet in 2007.[39] By the end of the investigation, approximately 1500 case subjects were identified in 43 states, the District of Columbia, and Canada, with the highest incidence rates occurring in New Mexico and Texas.

Cluster investigations were started on the cases, with questionnaires being utilized to locate commonalities of those potentially exposed. Salsa and guacamole, both foods typically containing tomatoes and hot peppers, were implicated repeatedly in cluster investigations, thus leading many to think tomatoes were the source of the *Salmonella*, since they have been implicated many times.[42-44] However, as the investigation proceeded, it was indicated that hot peppers, including jalapeño and Serrano peppers, were the vehicle for transmission. There is an amount of collinearity due to the fact that tomatoes and peppers are eaten together in things like salsa and guacamole.[39]

After an environmental investigation conducted by the FDA, the pepper contamination may have occurred on the farm, which may not be that unusual. A small survey of Mexican farms that grow chile peppers indicated that 6 of 14 irrigation-water samples (43%) and 3 of 5 pepper rinses (60%) yielded *salmonella*, although none of the serotypes were Saintpaul.[45] This outbreak investigation highlighted the challenges of epidemiologic identification of ingredients in foods that are commonly consumed, rapid identification and investigation of local clusters, the need to continue exploring hypotheses during an ongoing outbreak, and produce tracing in the supply chain.[39]

2.3b *SALMONELLA* TENNESSEE

Salmonella Tennessee infections are rare and most of the sources of infection are unknown.[46] An average of 52 cases were reported annually during 1995-2004, 0.1% of all *Salmonella* strains, but the only reported outbreak previously reported was due to contaminated powdered milk.[47,48] In November 2006, there was a widespread increase in the number of PFGE identified isolates of *Salmonella* Tennessee reported to PulseNet. By

December, 52 isolates from 25 states, and by July 2007, a total of 715 cases were identified in 48 states.[46]

In the initial investigation, it was noted that cases were not geographically clustered nor were there common food exposures when patient interviews were conducted locally. When the investigation was expanded using a standard food consumption survey, it was noted that of the 31 patients interviewed, 48% ate turkey and 85% consumed peanut butter.[46] Epidemiologic data suggested that two brands of peanut butter were the possible sources of the outbreak, both of which were manufactured in the same plant. This led to the company making radical changes at its manufacturing plant.

This was the first reported peanut butter outbreak reported in the US. Peanut butter was previously considered as low risk for *Salmonella* contamination, but that is no longer the case. Peanuts could have become contaminated with salmonellae during growth, harvest, or storage. Salmonellae can enter food-processing plants by various mechanisms, such as through raw agricultural products, water, animals, humans, or other surfaces.[46] The organisms are able to survive high temperatures in high-fat, low water activity environments.[49] Peanut butter provides such an environment, and *Salmonella* has been shown to survive for at least 6 months in peanut butter.[50] This outbreak helped to reveal the potential for widespread illness from a broadly distributed product with a long shelf life and that *Salmonella* surveillance with serotyping is critical in detecting such outbreaks. It also showed that processed foods have the potential to be contaminated in many steps, which shows the need for effective controls in food processing plants.

2.3c *SALMONELLA* TYPHIMURIUM

In November 2008, PulseNet detected a cluster of *Salmonella* Typhimurium with the same, rare PFGE pattern in 16 states and later in the month a second cluster of 27 isolates in 14 states with two rare PFGE patterns was noted.[51] Due to similarity, this was defined as an outbreak that identified 714 cases from 46 states. Of these cases, 86 hypothesis-generating interviews from 26 states noted that 47 of 81 respondents (58%) reported having been exposed to institutional settings; 56 of 79 (71%) reported eating peanut butter; and 61 of 71 (86%) reported eating chicken. However, respondents reported eating many different types and brands of peanut butter and chicken products.[51] Ongoing interviews revealed that peanut butter was the source of the outbreak, with one manufacturer at the heart of the problem.

This nationwide outbreak was linked to eating contaminated peanut butter, peanut paste, and roasted peanuts produced at the PCA facilities in Georgia and Texas. This outbreak resulted in one of the largest food recalls in U.S. history and an estimated $1 billion loss in peanut sales.[52] The traceback investigations led to multiple possibilities for the source of the contamination. *Salmonella* can survive in a low-moisture food such as peanut butter for at least 24 weeks; therefore, if postprocessing contamination occurs, *Salmonella* may survive in peanut butter for its entire shelf life of 18 to 24 months.[50] Contamination of low-moisture foods is likely to lead to prolonged, dispersed outbreaks that may be sustained as long as production conditions lead to contamination. The duration of the outbreak and range of production dates among *Salmonella*-positive food samples suggest that the outbreak strain may have been present in the PCA facilities for an extended period.[51] This outbreak was instrumental in refocusing national attention on food safety and spurring discussions about gaps in the food safety system and methods for establishing and enforcing basic preventive controls.

2.3d *SALMONELLA* NEWPORT

Salmonella Newport causes more than an estimated 100,000 infections annually in the United States, making it the third most common serotype causing human illness.[44] From July to November 2005, 72 laboratory confirmed *S.* Newport isolates indistinguishable by PFGE from the outbreak strain were identified in 16 states. During the investigation, it was found that most cases (70%) were exposed to uncooked tomatoes in restaurants than any other item. Of 27 cases, 11 (41%) reported eating beefsteak tomatoes, and 13 (48%) reported eating other types of tomatoes. Twenty-six (90%) of 29 cases and 86 (72%) of 119 controls had any exposure to tomatoes in either a home or restaurant.[44]

Traceback investigation showed that two growers/packing houses on the eastern shore of Virginia were the source of the contaminated tomatoes. Farms in this region supplied only the eastern and central United States at the time of this outbreak, matching the national distribution of cases of the outbreak pattern of *S.* Newport.[44] This outbreak was actually much larger than reported. Given that about one of every 38 cases of sporadic, laboratory-confirmed *Salmonella* infection is ascertained by public health surveillance, it was estimated that more than 2,500 patients might have been affected by this outbreak.[53] This outbreak advanced awareness of produce-associated outbreaks.

2.4 RANDOM FORESTS

Random forests is a machine-learning tool used for classification with applications in big data. Most uses of it applications in epidemiology are in genetic studies. Random forests classify by inputting a new object down each of the trees in the forest.[54] In a random forest, a number of decision trees are built during the process. Since there are many trees built in the process of running a random forest algorithm, it is called a forest. To classify a new object from an input variable, put said variable down each of the trees in the forest. It is a model that uses binary splits on independent variables to predict outcome, read like a flow chart. Random forests iteratively develops decision trees which can be used in categorical or continuous variable prediction.[54] Each tree classifies each observation into a particular category and the tree "votes" for that category. The forest chooses the category having the most votes over all the trees in the forest. The underlying algorithms are highly accurate, can run quickly on large databases, and can give estimates of what variables are important in classification, referred to as "variable importance". Random forests is an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.[54]

The core building block of random forests is a CART (classification and regression tree) inspired decision tree. The CART algorithm starts by drawing a random sample of individuals from the main dataset and building a decision tree based on this sample. Then, it repeats the process a second time, picking another random sample and growing a second decision tree. The prediction from the second tree will typically be different than those of the first tree.[55] This process continues, generating more trees each built on a slightly different sample and generating at least slightly different predictions each time. Random forests builds upon CART by adding randomness into the actual tree growing and not just the sampling.[54] Random forests takes a randomized sample of the rows in the dataset, creating a collection of unique trees which all make their classifications differently. Each tree is called to make a classification, the "votes" are tallied, and the majority decision is chosen. Since each tree is grown out fully, they each overfit, but in different ways. Thus, the mistakes one makes will be averaged out over them all.[55]

Random forests also result in a measure of variable importance. This method measures the relative importance of a variable correctly predicting the outcome category. It is based on measuring the damage that would be done to our predictive models if we lost access to true values of a given variable.[56] The more the accuracy of the random forest decreases due to the exclusion (or permutation) of a single variable, the more important that variable is deemed. Hence, variables with a large mean decrease in accuracy are more important for classification of the data.[57] While that measures accuracy, there is another measure, GINI. GINI is based on the actual role of a predictor and offers an alternative importance assessment based on the role the predictor plays in the data. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest.[57] Each time a particular variable is used to split a node, the Gini coefficient for the subsequent child nodes are calculated and compared to that of the original node. The Gini coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous). The changes in Gini are summed for each variable and normalized at the end of the calculation. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient.[57]

Currently, hierarchical cluster analysis is the main method of identifying similarities and differences among serotypes of *Salmonella*. This method results in clusters formed in a hierarchical fashion, which may be less efficient than using a method like random forests.[58] Most uses of random forests in a foodborne illness setting do not extend past looking at the PFGE patterns to determine similarities in serotypes, something that will be achieved here.[58,59] The importance of this work will be to attempt to use a method currently more focused on either genetic or microbiological studies and apply them to an epidemiological setting. This work will focus on finding a group of foods that will contain the true cause of an outbreak. This could result in faster and more accurate resolutions to outbreaks than the currently used case studies or hierarchical cluster analysis.

2.5 WORKS CITED

1. Nyachuba DG. Food-borne illness: is it on the rise? Nutr Rev. 2010; 68: 257-269.
2. Centers for Disease Control and Prevention (CDC). Surveillance for foodborne disease outbreaks—United States, 2007. *Morbidity and Mortality Weekly Report.* 2010;59(31):973–979.
3. Centers for Disease Control and Prevention. Foodborne Illness, Foodborne Disease, (sometimes called “Food Poisoning”). http://www.cdc.gov/foodsafety/facts.html#what. Published 2014. Accessed 1 April 2014.
4. National Institute of Diabetes and Digestive and Kidney Diseases. Foodborne Illnesses. http://digestive.niddk.nih.gov/ddiseases/pubs/bacteria/. Published 2012. Accessed 1 April 2014.
5. US Department of Health and Human Services. Food Poisoning. http://www.foodsafety.gov/poisoning/. Published 2014. Accessed 1 April 2014.
6. Santos RL, Raffatellu M, Bevins CL, Adams LG, Tükel C, Tsolis RM, Bäumler AJ. Life in the inflamed intestine, *Salmonella* style. Trends Microbiol. 2009:17, 498–506.
7. Griffin AJ, McSorley SJ. Development of protective immunity to *Salmonella*, a mucosal pathogen with a systemic agenda. Mucosal Immunol. 2011;4:371–382.
8. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ. Typhoid fever. *N. Engl. J. Med.*. 2002; 347: 1770-1782.
9. Centers for Disease Control and Prevention (CDC). Multistate outbreaks of *Salmonella* infections associated with raw tomatoes eaten in restaurants—United States, 2005–2006. *Morbidity and Mortality Weekly Report.* 2007;56:909-911.

10. Centers for Disease Control and Prevention (CDC). Outbreak of *Salmonella* serotype Saintpaul infections associated with multiple raw produce items—United States, 2008. *Morbidity and Mortality Weekly Report.* 2008;57(34):929-934.
11. Centers for Disease Control and Prevention (CDC). Multistate outbreak of *Salmonella* infections associated with peanut butter and peanut butter-containing products—United States, 2008–2009. *Morbidity and Mortality Weekly Report.* 2009;58:85-90.
12. Centers for Disease Control and Prevention (CDC). Multistate outbreak of *Salmonella* serotype Tennessee infections associated with peanut butter—United States, 2006-2007. *Morbidity and Mortality Weekly Report.* 2007;56(21):521-524.
13. Olsen SJ, MacKinnon LC, Goulding JS, Bean NH, Slutsker L. Surveillance for food-borne-disease outbreaks—United States, 1993–1997. *Morbidity and Mortality Weekly Report Surveillance Summary.* 2000;49:1-62.
14. Rabsch W, Tschäpe H, Bäumler AJ. Non-typhoidal salmonellosis: emerging problems. *Microbes Infect*. 2001; 3(3):237-247.
15. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV. Food-related illness and death in the United States. *Emerg. Infect Dis.* 1999; 5: 607-625.
16. Chalker RB, Blaser MJ. A review of human salmonellosis: III. Magnitude of *Salmonella* infection in the United States. *Rev. Infect. Dis.* 1988;10:111-124.
17. Glynn MK, Bopp C, Dewitt W, Dabney P, Mokhtar M, Angulo FJ. Emergence of multidrug-resistant Salmonella enterica serotype typhimurium DT104 infections in the United States, *N. Engl. J. Med.* 1998;338:1333–1338.
18. Angulo FJ, Swerdlow DL. *Salmonella* enteritidis infections in the United States, *J. Am. Vet. Med. Assoc.* 1998;213:1729–1731.
19. Cohen ML, Tauxe RV. Drug-resistant *salmonella* in the United States: an epidemiologic perspective, *Science.* 1986;234:964–969.
20. St. Louis ME, Morse DL, Potter ME, DeMelfi TM, Guzewich JJ, Tauxe RV, Blake PA. The emergence of grade A eggs as a major source of *Salmonella* enteritidis infections. New implications for the control of salmonellosis. *JAMA* 1988;259:2103–2107.
21. Rodrigue DC, Tauxe RV, Rowe B. International increase in *Salmonella* enteritidis: a new pandemic? *Epidemiol. Infect.* 1990;105:21–27.
22. Frenzen PD, Riggs TL. *Salmonella* cost updated using foodnet data. *Food Review*. 1999;22(2):10–15.
23. Walsh AL, Phiri AJ, Graham SM, Molyneux EM, Molyneux ME. Bacteremia in febrile Malawian children: clinical and microbiologic features. *Pediatr Infect Dis J.* 2000;19:312-318.
24. Molyneux EM, Walsh AL, Malenga G, Rogerson S, Molyneux ME. *Salmonella* meningitis in children in Blantyre, Malawi, 1996-1999. *Ann Trop Paediatr.* 2000;20:41-44.
25. Centers for Disease Control and Prevention. National Notifiable Diseases Surveillance System (NNDSS). http://wwwn.cdc.gov/nndss/default.aspx Published 2014. Accessed 10 March 2015.

26. Centers for Disease Control and Prevention. About FoodCORE. http://www.cdc.gov/foodcore/about.html. Published 2014. Accessed 1 April 2014.
27. Holmberg SD, Wachsmuth IK, Hickman-Brenner FW, Cohen ML. Comparison of plasmid profile analysis, phage typing, and antimicrobial susceptibility testing in characterizing *Salmonella* typhimurium isolates from outbreaks. *J Clin Microbiol.* 1984;19:100-4.
28. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg Infect Dis.* 2001;7: 382–389.
29. Ackers ML, Mahon BE, Leahy E, Goode B, Damrow T, Hayes PS, et al. An outbreak of *Escherichia coli* O157:H7 infections associated with leaf lettuce consumption. *J Infect Dis*. 1998;177:1588-93.
30. Barrett TJ. Molecular fingerprinting of foodborne pathogenic bacteria: An introduction to methods, uses and problems. In: Tortorello ML, Gendel SM, editors. Food microbiological analysis: new technologies. New York: Marcel Dekker; 1997. p. 249-64.
31. Graves LM, Swaminathan B, Hunter SB. Subtyping *Listeria monocytogenes*. In: Ryser EM, Marth EH, editors. *Listeria*, listeriosis and food safety. New York: Marcel Dekker; 1999. p. 279-98.
32. Jimenez A, Barros-Velazquez J, Rodriguez J, Villa TG. Restriction endonuclease analysis, DNA relatedness and phenotypic characterization of *Campylobacter jejuni* and *Campylobacter coli* isolates invovled in food-borne disease. *J Appl Microbiol.* 1997;82:713-21.
33. Maslanka SE, Kerr JG, Williams G, Barbaree JM, Carson LA, Miller JM, et al. Molecular subtyping of *Clostridium perfringens* by pulsed-field gel electrophoresis to facilitate food-borne-disease outbreak investigations. *J Clin Microbiol.* 1999;37:2209-14.
34. Threlfall EJ, Hampton MD, Ward LR, Rowe B. Application of pulsed-field gel electrophoresis to an international outbreak of *Salmonella agona*. *Emerg Infect Dis.* 1996;2:130-2.
35. Threlfall EJ, Ward LR, Hampton MD, Ridley AM, Rowe B, Roberts D, et al. Molecular fingerprinting defines a strain of *Salmonella enterica* serotype Anatum responsible for an international outbreak associated with formula-dried milk. *Epidemiol Infect*. 1998;121:289-93.
36. Wachsmuth K. Molecular epidemiology of bacteria infections: Examples of methodology and of investigations of outbreaks. *Rev Infect Dis.* 1986;8:682-92.
37. Barrett TJ, Lior H, Green JH, Khakhria R, Wells JG, Bell BP, et al. Laboratory investigation of a multi-state food-borne outbreak of Escherichia coli O157:H7 by using pulsed-field gel electrophoresis and phage typing. *J Clin Microbiol.* 1994;32:3013-7.
38. Stephenson J. New approaches for detecting and curtailing foodborne microbial infections. *JAMA.* 1997;277:1337-40.
39. Barton Behravesh C, Mody RK, Jungk J, et al for the *Salmonella* Saintpaul Outbreak Investigation Team. 2008 outbreak of *Salmonella* Saintpaul infections associated with raw produce. *N Engl J Med*. 2011; 364: 918-927.

40. Sivapalasingam S, Friedman CR, Co- hen L, Tauxe RV. Fresh produce: a growing cause of outbreaks of foodborne illness in the United States, 1973 through 1997. *J Food Prot*. 2004;67:2342-53.
41. Lynch MF, Tauxe RV, Hedberg CW. The growing burden of foodborne outbreaks due to contaminated fresh produce: risks and opportunities. *Epidemiol Infect.* 2009;137:307-15.
42. Hedberg CW, Angulo FJ, White KE, et al. Outbreaks of salmonellosis associated with eating uncooked tomatoes: implications for public health. *Epidemiol Infect.* 1999;122:385-93.
43. Gupta SK, Nalluswami K, Snider C, et al. Outbreak of *Salmonella* Braenderup infections associated with Roma tomatoes, northeastern United States, 2004: a useful method for subtyping exposures in field investigations. *Epidemiol Infect.* 2007;135: 1165-73.
44. Greene SK, Daly ER, Talbot EA, et al. Recurrent multistate outbreak of *Salmonella* Newport associated with tomatoes from contaminated fields, 2005. *Epidemiol Infect.* 2008;136:157-65.
45. Gallegos-Robles MA, Morales-Loredo A, Alvarez-Ojeda G, et al. Identification of *Salmonella* serotypes isolated from cantaloupe and chile pepper production systems in Mexico by PCR-restriction fragment length polymorphism. *J Food Prot.* 2008;71:2217-22.
46. Sheth AN, Hoekstra M, Patel N, Ewald G, Lord C, Clarke C, Villamil E, Niksich K, Bopp C, Nguyen T, Zink D, Lynch M. A national outbreak of *Salmonella* serotype Tennessee infections from contaminated peanut butter: a new food vehicle for salmonellosis in the United States. *Clin Infect Dis.* 2001;53: 356–362.
47. Centers for Disease Control and Prevention. *Salmonella* surveillance summary, 2004. http://www.cdc.gov/ncidod/dbmd/phlisdata/salmonella.htm. Published 2006. Accessed 1 April 2014.
48. Centers for Disease Control and Prevention (CDC). *Salmonella* serotype Tennessee in powdered milk products and infant formula—Canada and United States, 1993. *Morbidity and Mortality Weekly Report.* 1993; 42:516–7.
49. Mattick KL, Jorgensen F, Legan JD, Lappin-Scott HM, Humphrey TJ. Habituation of *Salmonella* spp. at reduced water activity and its effect on heat tolerance. *Appl Environ Microbiol*. 2000; 66:4921–5.
50. Burnett SL, Gehm ER, Weissinger WR, Beuchat LR. Survival of *Salmonella* in peanut butter and peanut butter spread. *J Appl Microbiol.* 2000; 89:472–7.
51. Cavallaro E, Date K, Medus C, et al. *Salmonella* Typhimurium infections associated with peanut products. *New Engl J Med.* 2011; 365: 601‑610.
52. ThomasNet News. Peanut recall sparks large-scale food safety concerns. http://news.thomasnet.com/IMT/archives/2009/03/salmonella-related-peanut-recalls-impact-on-manufacturers-businesses-becoming-clearer.html. Published 2009. Accessed 1 April 2014.
53. Voetsch AC, Van Gilder TJ, Angula FJ, et al. FoodNet estimate of the burden of illness caused by nontyphoidal *Salmonella* infections in the United States. *Clin Infect Dis.* 2004; 38: S127–134.

54. UC Berkeley. Random Forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. Published 2001. Accessed 20 Sep 2014.
55. Stephens T. Titanic: Getting Started With R - Part 5: Random Forests. http://trevorstephens.com/kaggle-titanic-tutorial/r-part-5-random-forests/. Published 2014. Accessed 28 Feb 2017.
56. Salford Systems. Random Forests for Beginners. http://info.salford-systems.com/an-introduction-to-random-forests-for-beginners. Published 2014. Accessed 19 Feb 2017.
57. Dinsdale et al. Random Forests. https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html. Published 2016. Accessed 31 Mar 2017.
58. Zou W, Lin WJ, Foley SL, Chen CH, Nayak R, Chen JJ. 2010. Evaluation of pulsed-field gel electrophoresis profiles for identification of *Salmonella* serotypes. *J. Clin. Microbiol*. 48:3122–3126.
59. Zou W, Lin WJ, Hise KB, Chen HC, Keys C, Chen JJ. 2012. Prediction system for rapid identification of *Salmonella* serotypes based on pulsed-field gel electrophoresis fingerprints. *J. Clin. Microbiol*. 50:1524–1532.

CHAPTER 3

METHODS

3.1 STUDY DESIGN

A case-only study design was used for Aim 1, and a cross-sectional study design was used for Aims 2 and 3. The dataset used for Aim 1 and Aim 2 was collected from standardized surveillance questionnaires submitted to the South Carolina Department of Health and Environmental Control (DHEC). The following events occurred for an illness to be ascertained and thus for inclusion in the study (Aims 1 and 2). The ill person sought medical care and had a specimen that tested positive for Salmonella was submitted to DHEC Bureau of Laboratories (BoL) for testing. The laboratory test identified the causative agent and reported the illness to public health authorities, in this case DHEC Division of Acute Disease Epidemiology (DADE). DADE administered the appropriate questionnaire and the person who was ill completed it fully. Data were collected for this study from January 2008 to June 2015, with 4058 possible *Salmonella* cases identified for inclusion based on the above criteria. Through the questionnaire, information on the foods consumed in the past seven days was collected. Both Aim 1 and Aim 2 also used serological data as recorded by BoL. All participants have their serological data linked to their questionnaire at DADE. Participants were grouped by serotype as well as whether or not they were part of an outbreak.

For Aim 1 we further split the dataset down to three separate outbreaks. DHEC assigned outbreak status in this dataset based on their knowledge of sporadically occurring cases. They limited it by geographical location, serotype, date, and foods eaten, taking into account seasonality and a threshold of commonly occurring cases. Based on the definition provided, three known DHEC outbreaks were chosen for analysis, May 2015, August 2012, and May 2012. The May 2015 outbreak was suspected to be caused by fast food chain chicken with 24 suspected cases. The August 2012 outbreak had 17 suspected cases was thought to be caused by boiled peanuts, contact with an amphibian, or a waterborne exposure. The May 2012 outbreak was caused by guacamole, chips, and sour cream with 9 cases.

For Aim 3 we used a different dataset that assessed food preferences of South Carolinians. The South Carolina Department of Health and Environmental Control (DHEC) commissioned the University of South Carolina's Institute for Public Service and Policy Research (IPSPR) to conduct a telephone survey with a representative sample of the state's population two years of age or older. The questionnaire used in this survey was initially designed by DHEC staff and based on the hypothesis-generating questionnaire that is used by DHEC when it encounters cases of *Salmonella* and Shiga toxin-producing *E. coli (*STEC). The modified questionnaire became the Food Exposures Survey, used to address the questions of foods eaten without being in relation to an outbreak or confirmed case of foodborne illness. This helps to aid in the assessment of overall eating patterns and not just those around the time the study participant may have gotten ill. The survey was administered to a target population aimed at being a representative sample of the entire state. There were 875 participants included in this analysis. Data collection was done via telephone, both cell phone and landline, from July 27-August 30, 2012 from 9:00 AM to 9:30 PM Monday through Friday, from 10:00 AM to 4:00 PM on Saturday, and 3:00 PM to 8:00 PM on Sunday. Table 3.1 shows a summary of datasets used.

Outbreaks were identified in multiple ways for the purposes of this research. First, the DHEC outbreak definition as given in the data. DHEC has more knowledge of sporadically occurring cases and thus, they are more stringent with their outbreak definition. They identify outbreaks by serotype, date, and foods eaten. DHEC defines an outbreak as two or more persons not living in the same household with the same enteric illness following a common exposure. Secondly, the FDA/CDC definition was used. They both define an outbreak as two or more cases of foodborne illness occurring during a limited period of time, here 30 days, with the same organism that are associated with either the same food service operation, such as a restaurant, or the same food product.[1,2] From here, it will be referred to as the FDA definition. The DHEC definition was used for Aim 1, while for Aim 2 we used both the DHEC and FDA definitions. Table 3.2 provides a summary of what outbreak was used in each case.

TABLE 3.1 SUMMARY OF DATASETS USED

	Population	Time	Aim 1	Aim 2	Aim 3
Serological Data	4058	7.5 years	X	X	
Hypothesis-Generating Questionnaire	4058	7.5 years	X	X	
Food Exposure Survey	875	1 month			X

TABLE 3.2 SUMMARY OF OUTBREAK DEFINITIONS USED

	Aim 1	Aim 2	Aim 3
DHEC	X	X	
FDA		X	

3.2 STUDY POPULATION

Setting Aims 1 and 2: All South Carolinians who went to see a health care provider for symptoms of foodborne illness, submitted a stool sample for *Salmonella* testing, and received a confirmation of illness were eligible to fill out the state-mandated hypothesis-generating questionnaire on foodborne illness (See appendix B). DHEC staff attempted to call all eligible participants, but not all were administered the questionnaire as some were lost to follow-up and others did not fully complete the questionnaire. All completed questionnaires and serotypes were collected from eighteen counties (Pickens, Greenville, Spartanburg, Cherokee, Union, Richland, Lexington, Newberry, Fairfield, Chester, Lancaster, York, Horry, Williamsburg, Georgetown, Berkeley, Dorchester, Charleston). The counties were selected for this sentinel project based upon: 1) disease burden (66% of all cases of *Salmonella*, Shiga toxin-producing Escherichia coli (STEC) and *Listeria*); 2) geographic diversity (Upstate, Midlands and Coastal areas); 3) population density (65% of the total state population); and, 4) presence of three metropolitan statistical areas (MSAs) listed in the "Top 100" nationally within the selected counties.

Time Period Aims 1 and 2: All individuals who participated in the study and filled out a questionnaire were included in the analysis. Data for this study was collected from January 2008 to June 2015, which is a total of 7.5 years.

Setting Aim 3: The study population of this aim was a random sample of South Carolina residents two years and older. A dual sampling frame approach was used in selecting study participants. One sampling set was based on landline telephone exchanges and the second on cell phone telephone numbers. For the landline component, respondents were selected from a random sample of households with telephones in the state. Respondents in the cell phone sample were randomly selected from a list of cell phone exchanges in South Carolina. All phones with a South Carolina exchange were eligible for the study and all counties were included in the sample. The survey interviewers called each of these numbers. Numbers that were found to be businesses, institutions, not in service, or otherwise not assigned were ineligible for the survey.

The remaining non-excluded numbers were called, which resulted in contact in both the landline component and the cell phone component of the study. When contact was made with a residence in the landline component, a participant two years of age or older was randomly chosen from the occupants of the household. If the selected participant was between the ages of 2 and 11, an adult in the household was asked to be a proxy to answer the questions for the child. Participants aged 12 or older were interviewed directly about their food exposure experiences. Proxy interviews were conducted for participants between the ages of 12 and 17 if an adult in the household did not want the selected child to participate. The food preferences questionnaire used in this survey was based on the hypothesis-generating questionnaire that is used by DHEC when it encounters a confirmed case of *Salmonella* or STEC (See Appendix C).

Time Period Aim 3: Data collection was done via telephone from July 27-August 30, 2012 from 9:00 AM to 9:30 PM Monday through Friday, from 10:00 AM to 4:00 PM on Saturday, and 3:00 PM to 8:00 PM on Sunday.

3.3 DATA ANALYSIS

AIM 1: In order to determine food(s) that may have caused a foodborne outbreak, investigators used the questionnaire data obtained from DHEC and serotype data obtained by DHEC BoL and identified those individuals associated with an outbreak by grouping according to the serotype data. DHEC staff attempted to call all South Carolinians who went to see a medical physician for symptoms of foodborne illness, submitted a stool sample for *Salmonella* testing, and received a confirmation of illness were eligible to fill out the hypothesis-generating questionnaire on foodborne illness (See appendix B). Not all eligible participants were administered the questionnaire as some were lost to follow-up and others did not fully complete the questionnaire.

The following events had to have occurred for an illness to be ascertained and thus, for inclusion into the study. The ill person sought medical care and a specimen was submitted to SC DHEC Bureau of Laboratories (BoL) for testing. For the purposes of this analysis, only *Salmonella* cases were considered. The laboratory test identified the causative agent as *Salmonella* and reported the illness to SC DHEC Division of Acute Disease Epidemiology (DADE). DADE then administered the appropriate questionnaire and the person who was ill completed it fully, stating what had been eaten in the seven days prior to illness onset. All completed questionnaires and serotypes were collected from eighteen counties (Pickens, Greenville, Spartanburg, Cherokee, Union, Richland, Lexington, Newberry, Fairfield, Chester, Lancaster, York, Horry, Williamsburg, Georgetown, Berkeley, Dorchester, Charleston). The counties were selected for this sentinel project based upon: 1) disease burden (66% of all cases of *Salmonella*, Shiga toxin-producing Escherichia coli (STEC) and *Listeria*); 2) geographic diversity (Upstate, Midlands and Coastal areas); 3) population density (65% of the total state population); and, 4) presence of three metropolitan statistical areas (MSAs) listed in the “Top 100” nationally within the selected counties.

Serotypes were grouped by dates of illness onset, serotype pattern, and whether or not it was a confirmed case as determined by DHEC BoL. Once grouped, the dataset was narrowed down to three separate outbreaks. DHEC assigned outbreak status in this dataset based on their knowledge of sporadically occurring cases. They limited it by geographical location, serotype, date, and foods eaten, taking into account seasonality and a threshold of commonly occurring cases. Based on the definition provided, three known DHEC outbreaks were chosen for analysis, May 2015, August 2012, and May 2012. The May 2015 outbreak was suspected to be caused by fast food chain chicken with 24 suspected cases. The August 2012 outbreak had 17 suspected cases was thought to be caused by boiled peanuts, contact with an amphibian, or a waterborne exposure. Guacamole, chips, and sour cream with 9 cases caused the May 2012 outbreak.

Data from each outbreak was run through the random forests package in R (randomForest). The package was run to determine a shortened list of food or environmental exposures responsible. The random forests algorithm drew a random sample from the main dataset and built a decision tree based on this sample. The package repeated multiple times, each time picking another random subset of data and growing a decision tree for each random subset. The prediction from each tree will typically be different than those of the other trees.[3]

The forest grown was used to calculate variable importance, a method to measure the relative importance of any variable. First, there is mean decrease accuracy. The more the accuracy of the random forest decreases due to the exclusion of a single variable, the more important that variable is deemed. Random forests performs this action one variable at a time to aid in the measurement of the loss of accuracy. Thus, variables with a large mean decrease in accuracy are more important for classification of the data.[3] There is also another variable importance measure, GINI. The mean decrease in GINI is a measure of how each variable contributes to the homogeneity of the resulting random forest. Each time a particular variable is used to split a note, the GINI coefficient for each of the subsequent nodes are calculated and compared to the original node. GINI is measured from 0 to 1.[4] The changes in the GINI coefficient are summed for each variable and normalized at the end of the random forest calculation. Variables with higher purity have a higher decrease in GINI.

From here, using the variable importance measures (mean decrease accuracy and GINI) a comparison on the effectiveness of random forests to traditional methods was completed. Food and environmental exposures for past outbreaks was collected from DHEC for comparison against the list of exposures generated from random forests. This comparison led to more defined methods of classification for the foods responsible for foodborne illness. Data management and analysis was done using R and SAS 9.4 software.

AIM 2: In order to determine dietary intake of South Carolinians involved in a foodborne outbreak, investigators used the same dataset as AIM 1, collecting data from January 2008 to June 2015, with 4058 possible *Salmonella* cases identified for inclusion. It was then limited down to those individuals associated with an outbreak by grouping according to the serotype data. This was done using two separate outbreak definitions. The FDA defines an outbreak as two or more cases of foodborne illness caused by the same organism that occur within a limited period of time and are associated with either the same food or same food service operation.[1] This study interpreted this as two or more cases of the same organism that occurred within 30 days and used this to find the food commonalities. Since DHEC has more knowledge of sporadically occurring cases, they are more stringent with their outbreak definition and limit it by serotype, date, and foods eaten. This definition has been provided to us in the data collected from DHEC. Of the 4058 *Salmonella* cases identified by DHEC from 2008-2015, DHEC identified 78 as being part of a statewide outbreak while using the FDA definition yielded 2565 cases as part of an outbreak. Dietary patterns were assessed using the same data in AIM 1 in conjunction with guidance available from the Center for Nutrition Policy and Promotion.[5] There were 186 food exposures categorized for analysis. The groups were poultry (8), meat (18), pork (6), seafood (9), egg (3), dairy (24), fruit (35), vegetables (26), greens (19), snacks (17), nuts (18), grains (3), raw foods (14), frozen foods (18), and prepackaged foods that are not frozen (17). The raw, frozen, and prepackaged categories contain foods that are already in the main categories. Each group was categorized as 0 for no and 1 for yes.

The investigators also looked at gender, racial, ethnic, and geographical differences among those identified in an outbreak. Urban or rural was categorized by using the county of residence from the questionnaire and then the rural definition based on the Office of Management and Budget (OMB) metropolitan statistical areas. A metro area includes one or more counties containing a core urban area of 50,000 or more people, together with any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core. OMB also defines micropolitan statistical areas using the same method but centered on urban areas with at least 10,000 but no more than 50,000 people.[6]

Gender was classified as 0 for female and 1 for male for ease of analysis. Age was categorized as an integer with the option of being missing if the participant refused. Race was made into a numeric variable and put into categories of white, African American, other, and unknown/refused. Hispanic was asked as whether or not the person being interviewed identified as Hispanic with the answers being categorized as 0 for no, 1 for yes, 99 and missing as unknown. The model used for analysis was a log-linear regression model with a Poisson distribution. Forward, backward, and stepwise selection methods were used to find the best variable selection. Relative risks were also calculated for each food exposure group and demographic variable. Data management and statistical analyses were performed using SAS 9.4 software.

AIM 3: Aim 3 conducted similar analyses to AIM 2, but used a different set of data. The investigators sought to characterize the nutrition and dietary intake of South Carolina residents using the South Carolina Department of Health and Environmental Control Food Exposures Survey. The questionnaire used in this survey was initially designed by DHEC staff and based on the hypothesis-generating questionnaire that is used by DHEC when it encounters cases of *Salmonella* and Shiga toxin-producing *E. coli* (STEC). The modified questionnaire became the Food Exposures Survey, used to address the questions of foods eaten without being in relation to an outbreak or confirmed case of foodborne illness. This helps to aid in the assessment of overall eating patterns and not just those around the time the study participant may have gotten ill. The survey was administered to a target population aimed at being a representative sample of the entire state. There were 875 participants included in this analysis. Data collection was done via telephone, both cell phone and landline, from July 27-August 30, 2012 from 9:00 AM to 9:30 PM Monday through Friday, from 10:00 AM to 4:00 PM on Saturday, and 3:00 PM to 8:00 PM on Sunday. Demographic, social, and geographic determinants of dietary patterns as identified by the guidance available from the Center for Nutrition Policy and Promotion was used in the analysis.[5]

There were 154 food exposures categorized for analysis. The groups were poultry (7), meat (10), pork (7), seafood (8), egg (2), dairy (20), fruit (29), vegetables (25), greens (13), snacks (15), nuts (15), grains (3), raw foods (26), frozen foods (15), and prepackaged foods that are not frozen (17). The raw, frozen, and prepackaged categories contain foods that are already in the main categories. Each group was categorized as 0 for no and 1 for yes. Other variables used in analysis included demographic groups broken down as urban or rural environment, age categories, gender, Hispanic ethnicity, income, and race. Urban or rural was categorized by using the county of residence from the questionnaire and then the rural definition based on the Office of Management and Budget (OMB) metropolitan statistical areas and their categorization of each county in the state of South Carolina.[6]

Gender was classified as 0 for female and 1 for male for ease of analysis. Age was categorized as an integer with the option of being missing if the participant refused. Race was put into categories of Caucasian, African American, other, and unknown/refused. Hispanic was asked as whether or not the person being interviewed identified as Hispanic with the answers being categorized as 0 for no, 1 for yes, 88 for refused, 99 and missing as unknown. Income was classified into categories of <$25,000, $25-49,999, $50-99,999, ≥

$100,000, and blank for refused or missing. For calculations of relative risk, each demographic was made into binary variables. Gender, Hispanic, and urban remained as they were with female/male, no/yes, and urban/rural, respectively. Age changed to < 45 and ≥ 45 years while race changed to Caucasian and other. Income became under $50,000 and $50,000 or greater.

The results were adjusted for age, race, and gender, using relative risks from Poisson regression for categorical variables in the comparison. Stepwise selection was also used for variable inclusion in the model. This study also determined if there was an association between demographic characteristics and the frequency of eating pre-packaged food or fresh foods. This was done using frequency and summary procedures as well as Poisson regression in SAS 9.4. Descriptive analyses were also performed.

3.4 SAMPLE SIZE

Aim 1 and Aim 2 used a dataset that contained 4058 possible *Salmonella* cases. Aim 1 then grouped serotypes by dates of illness onset, serotype pattern, and whether or not it was a confirmed case as determined by DHEC BoL. Once grouped, the dataset was narrowed down to three separate outbreaks. DHEC assigned outbreak status in this dataset based on their knowledge of sporadically occurring cases. They limited it by geographical location, serotype, date, and foods eaten, taking into account seasonality and a threshold of commonly occurring cases. The May 2015 outbreak was suspected to be caused by fast food chain chicken with 24 suspected cases. The August 2012 outbreak had 17 suspected cases was thought to be caused by boiled peanuts, contact with an amphibian, or a waterborne exposure. Guacamole, chips, and sour cream with 9 cases caused the May 2012 outbreak.

Aim 2 limited the original dataset to those individuals associated with an outbreak by grouping according to the serotype data. This was done using two separate outbreak definitions. The FDA defines an outbreak as two or more cases of foodborne illness caused by the same organism that occur within a limited period of time and are associated with either the same food or same food service operation.[1] This study interpreted this as two or more cases of the same organism that occurred within 30 days and used this to find the food commonalities. Since DHEC has more knowledge of sporadically occurring cases, they are more stringent with their outbreak definition and limit it by serotype, date, and foods eaten. This definition has been provided to us in the data collected from DHEC. Of the 4058 *Salmonella* cases identified by DHEC from 2008-2015, DHEC identified 78 as being part of a statewide outbreak while using the FDA definition yielded 2565 cases as part of an outbreak.

Aim 3 conducted similar analyses to AIM 2, but used a different set of data that originated from the South Carolina Department of Health and Environmental Control Food Exposures Survey. The survey was administered to a target population aimed at being a representative sample of the entire state. There were 875 participants included in this analysis. Data collection was done via telephone, both cell phone and landline, from July 27-August 30, 2012 from 9:00 AM to 9:30 PM Monday through Friday, from 10:00 AM to 4:00 PM on Saturday, and 3:00 PM to 8:00 PM on Sunday.

3.5 PROTECTION MEASURES

All study personnel were trained and certified in federal and state policies regarding the protection of human subjects' participation in research. The human subjects data used in AIMS 1, 2, and 3 of this dissertation were part of the research conducted by the South Carolina Department of Health and Environmental Control (DHEC). Careful consideration was taken to ensure the anonymity of study participants. No individual was in any publications resulting from this study. To further protect patient confidentiality, DHEC only supplied an ID number to identify the study participant, and thus investigators did not have any name information. The abstracted data from each questionnaire was entered into an electronic database which was used in the analysis. This investigation posed only a minimal risk to the privacy of individuals, and the research conducted fell under the non-Human Subjects Exemption (AIM 1, AIM 2, and AIM 3). This was because the specimens and/or private information/data were not collected specifically for the dissertation research project through an interaction/intervention with living individuals, AND the investigator(s) including collaborators on the proposed research could not readily ascertain the identity of the individuals(s) to whom the coded private information or specimens pertain. An application was submitted and approved by both the University of South Carolina Institutional Review Board and the DHEC Institutional review board (Please see Appendix A for a more detailed description of the Human Subjects Protection).

3.6 STRENGTHS AND LIMITATIONS

There is a strong need for quicker and more effective ways of identifying food(s) responsible for a foodborne outbreak. Most studies currently use hierarchical cluster analyses or case-only studies to parse through questionnaire information. Random Forests can help identify the food associated with a foodborne outbreak quicker. Advantages of random forests include the ability to analyze a large number of variables input into the model without variable deletion. This was especially important in our model since any one of the 207 exposures could potentially be the cause of the outbreak. Another advantage of using random forests is that its effectiveness is not limited by small cell sizes (unlike logistic regression and other parametric methods). For example, it was possible to apply the method to outbreaks that consisted of samples of 24 cases (May 2015), 17 cases (August 2012), and 9 cases (May 2012). The model ran efficiently even with the limited number of cases. Ultimately, the use of random forests could aid in reducing the number of people who get sick from foodborne pathogens.

Aim 1 of this study described the application of a novel method to identify foods associated with an outbreak more efficiently. Aim 2, described foods that were likely to be associated with *Salmonella* outbreaks in South Carolina among people investigated for outbreaks. This information would be helpful for future disease surveillance activities. Because participants were identified and surveyed immediately, since the state mandates that probable *Salmonella* cases be identified within three business days, it helped to avoid possible information bias. It has been shown that foods eaten in a particular week are correlated with what is eaten in the past.[7] If the ill person ate something out of character, it is likely they would remember since it possibly led to their illness.

The chances of recall bias were increased when the questionnaire was not be administered in a timely fashion, such as when the results of the laboratory test were not reported late. Finally, missing data or low response could cause selection bias. However, as all individuals participating in the survey had been diagnosed with Salmonella the errors would be evenly distributed between the outbreak and non-outbreak groups, which would likely cause non-differential misclassification.

A limitation of this study is that while random forests cannot exactly predict the true exposure that caused the outbreak. This could partially be due to randomness in the model, which could result in a random variable having a stronger value of importance than the true outbreak exposure. Our results did put the actual outbreak cause toward the top of the exposure list, but it was not the top answer. While it may not be possible to exactly identify the food associated with the outbreak using Random Forests, it is possible to shorten the list of likely foods causing it. In this study, we were able to reduce the list of possible exposures causing the outbreak from 207 to 30 within a matter of minutes. This step can potentially help DHEC to have a more focused follow-up investigation with the shortened list to identify the food causing the outbreak sooner and containing the outbreak.

Aim 3 of this study used the Food Exposures Survey and was administered to a representative sample of South Carolina residents over a one month period in the summer of 2012 with a questionnaire that was based on the hypothesis-generating questionnaire used by DHEC to investigate foodborne disease outbreaks. This questionnaire had 2 versions, version 1 with 145 questions and version 2 with 147 questions (see Appendix C). Each food exposure was categorized similarly to the hypothesis-generating questionnaire as yes/no/don't know for foods eaten in the past 7 days. However, the questionnaire was changed given the goal of limiting the time required to complete to 20 minutes and some questions were dropped.

When the survey was conducted, both landlines and cell phone exchanges in South Carolina were used. There is no issue with the landline as you must have a South Carolina area code to have a phone in your house, but issues can exist with the cell phone exchanges chosen. A representative sample of the South Carolina population may not be achieved since only SC area codes were also chosen for cell phones. If someone moved to South Carolina, such as a student moving here for college, they may not have chosen to change their cell phone number and thus have an exchange from another state. This removes a part of the population from being eligible for this study before any participants are chosen. To remediate this problem, the census weight was used in all calculations to create a more representative target population.

This survey, like all surveys, has the potential for sampling error given that only a part of the population of the state was interviewed and not all residents of South Carolina participated. For the questions that were answered by at least 800 respondents, the potential error is very low, but those answered by significantly less than 800 respondents have the potential for a larger variation than those for the entire sample.[8] To help reduce this potential error, foods and demographics were grouped to reduce the variation.

Another potential issue with this study could arise from the way the outbreak groups were defined. The DHEC definition of an outbreak takes into account what is going on across the country as well as what is known about South Carolina in particular. For example, we may see a rise in cases in the summertime due to family picnics and people leaving food out on a table, causing it to warm in the sun and spoil. This most likely is not cause for an outbreak and thus DHEC would not identify it as such. Due to this, there is a small number of cases (78) considered part of an outbreak using this definition. The very large number of potential cases identified by using the FDA definition suggests that it included a large number of false positive cases because foodborne outbreaks are not that common. The DHEC outbreak definition may therefore be more accurate even though it cannot be directly determined from these data.

The criteria for FDA definition consisted of the primary pattern of the organism and the illness onset date. If 2 or more cases occurred in a calendar month, then they were identified as being part of an outbreak. The problem with this definition is that it does not look at cases ± 30 days, but rather calendar month. So, if a case occurs on May 31st and another on June 2nd, they would not be identified as part of the same outbreak. This could lead to an underestimate of cases in the outbreak, but after combing through the data, cases that should have been a "yes" to being included in the outbreak definition were and this did not lead to any misclassification.

3.7 SUMMARY

This research is significant because we applied a new approach, random forests, which could shorten the time to identify foods associated with foodborne disease outbreaks. In addition, we characterized people who were part of a disease outbreak in South Carolina, and created a profile of South Carolina residents who consumed foods that put them at risk of foodborne diseases. We used data that was part of foodborne disease surveillance activities conducted by DHEC. Random Forests is a machine learning tool that is used to classify a large number of variables into smaller categories. We used Random Forests to shorten a long list of foods obtained from people who were part of an outbreak investigation, into a shorter list of foods that were likely to be associated with a foodborne disease outbreak. This step could help to identify the food causing an outbreak quicker. To the best of our knowledge, random forests has not been used for this purpose before. With foodborne outbreaks constituting a major, ongoing public health burden, prompt and effective detection of the source through epidemiologic investigations are necessary to remove contaminated food from the market, prevent further illnesses, and focus prevention strategies on critical contamination points. The research performed in conjunction with South Carolina's Department of Environmental Health and Control (DHEC) worked to enhance detection, investigation, and control of foodborne disease. There is needed research in the field of nutrition and foodborne outbreaks to determine any relationship as foodborne illness continues to become more prevalent. This information will address the challenges of a rural southern state with high disease burden in locations with the highest number of cases, something that has not previously been shown.

While foodborne pathogens and the study of outbreaks themselves are not innovative, the ways they are being investigated continue to evolve. Random Forests is a relatively new data driven machine-learning tool to identify predictive patterns in big data that is used in many diverse fields. Its application to foodborne disease outbreak investigations can potentially help to identify foods causing illness quickly, but to the best of our knowledge, this has not yet been done. We have applied Random Forests to identify foods associated with *Salmonella* outbreaks in South Carolina using outbreak investigation data collected by DHEC (Aim 1). The results of this study can potentially improve foodborne disease outbreak investigations in South Carolina and beyond.

In addition, we studied dietary patterns derived from questionnaires that DHEC administered to different groups of people in South Carolina. The first dietary pattern was derived from a food exposure questionnaire that had been administered to individuals as part of outbreak investigations. Food groups associated with *Salmonella* outbreaks were identified, and then described by demographic and other characteristics (Aim 2). The second food group pattern was derived from a questionnaire that was similar to the one DHEC had used for the outbreak investigation, but was administered to a representative sample of individuals living in South Carolina. We then described the demographic, social and geographic predictors of consumption of these food groups by South Carolina residents (Aim 3). This information characterizing food exposure data collected by DHEC as part of its surveillance will help in interpreting data collected in outbreak investigations. In summary, the analysis performed can help make identification of foods responsible for outbreaks faster and more efficient.

3.8 WORKS CITED

1. U.S. Food and Drug Administration. What You Should Know About Government Response to Foodborne Illness Outbreaks. http://www.fda.gov/Food/ResourcesForYou/Consumers/ucm180323.htm. Published 2014. Accessed 20 October 2014.
2. Centers for Disease Control and Prevention. Foodborne Disease Outbreak 2011 Case Definition. https://wwwn.cdc.gov/nndss/conditions/foodborne-disease-outbreak/case-definition/2011/. Published 2016. Accessed 31 Oct 2016.
3. Salford Systems. Random Forests for Beginners. http://info.salford-systems.com/an-introduction-to-random-forests-for-beginners. Published 2014. Accessed 19 Feb 2017.
4. Strickland J. "9. Decision Tree Learning." Data Analytics Using Open-Source Tools. N.p.: Lulu.com, 2016. 201.
5. US Department of Agriculture and US Department of Health and Human Services. Dietary Guidelines for Americans, 2010. 7th ed. Washington, DC: US Government Printing Office; 2010.
6. U.S. Department of Agriculture Economic Research Service. South Carolina - Rural Definitions: State Level Maps. http://www.ers.usda.gov/datafiles/Rural_Definitions/StateLevel_Maps/SC.pdf. Published 2015. Accessed 15 Nov 2015.
7. Rimm ED, Giovannucci EL, Stampfer MJ, Colitz GC, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. Am J Epidemiol . 1992;135: 1114-1126.
8. Oldendick, B. (2013). South Carolina Department of Health and Environmental Control Food Exposures Survey – Technical Report. South Carolina: University of South Carolina's Institute for Public Service and Policy Research.

CHAPTER 4

USE OF RANDOM FORESTS TO ESTIMATE FOOD AND ENVIRONMENTAL CAUSES OF *SALMONELLA* OUTBREAKS IN SOUTH CAROLINA[1]

[1] Alianell, A.T., Merchant, A., McLain, A., Brenner, E., and D. Giurgiutiu. To be submitted to *American Journal of Epidemiology*.

4.1 ABSTRACT

Foodborne illness remains a serious public health problem in the United States in general as well as South Carolina in particular. Obtaining good food ingestion histories as well as possible risky environmental exposures is one of the earliest, most important tasks to complete in any foodborne outbreak investigations. Because time is of the essence in investigations, we have evaluated a rarely used biostatistical method, *Random Forests,* that has the potential to facilitate more rapid identification of foods or environmental exposures that may be associated with outbreaks. We applied *Random Forests* to data provided by the South Carolina Department of Health and Environmental Control (DHEC), concerning three outbreaks of salmonellosis: (1) May 2012 (9 cases due to contaminated guacamole/chips/sour cream); (2) August 2012 (17 cases due to boiled peanuts/amphibian exposure); and (3) May 2015 (24 cases due to contaminated chicken). In each case, *Random Forests* helped pare down a list of over 200 potential food and environmental exposures to a much shorter list of just 30, each of which contained the eventually confirmed "cause" of the outbreak. We suggest that *Random Forests* may be more efficient and effective than current methods of investigating foodborne outbreaks, especially when the number of exposures potentially implicated may be large.

4.2 INTRODUCTION

Each year, *Salmonella* is estimated to cause one million foodborne illnesses in the United States, with 19,000 hospitalizations and 380 deaths.[1] Although foodborne illnesses affect all people, they are more likely to occur in infants and children, pregnant women, older adults, and people with weak immune systems. Most healthy individuals recover without treatment in about 4 to 7 days without the need for hospitalization.[2,3] To prevent foodborne illness from affecting many people, there is a need to quickly identify the specific food causing the outbreak. However, finding the source of contamination is challenging since contamination can occur anywhere along the food production chain. Pathways of enteric diseases like *Salmonella* are multifaceted due in part to the fact that sources of the disease may or may not be foodborne. Causality can change based on geographic location and demographics of the population, among other things.

A first step in the investigation of food borne disease outbreaks is to formulate questionnaires to ascertain many food and environmental exposures, sometimes containing hundreds of questions, like the South Carolina Department of Health and Environmental Control's (DHEC) hypothesis generating questionnaire.[4] Exposures collected from such questionnaires are structured in a way to obtain broad categories of exposure (i.e. consumption of breaded chicken in the last week) with yes/no/don't know responses. With such limitations, using conventional modeling like logistic regression can be difficult due to the fact that we eat many categories of food in a single meal.[5] Another issue that may arise is missing or non-response data, which may bias the estimation of the exposure-outcome relationship.[6] Due to the fact that exposures can overlap with multiple foods eaten in one meal as well as environmental exposures affecting the causality, new methods are needed to analyze such complex relationships.

Random forests is a method that has been successfully used to analyze complex datasets in biomedical studies.[7,8] It is a powerful machine learning tool used to classify many variables into groups based on defined data characteristics. Random forests starts with a standard machine learning technique, a decision tree. An input is entered at the top gets bucketed into smaller and smaller trees.[9] Random forests can capture non-linear relationships and interactions from data, which can be useful in studies with more than 20 variables of interest.[10] This is especially relevant in foodborne studies where there are many variables that can qualify as exposures. With conventional methods, it would be difficult to specify a model with all relevant exposures and the interactions they may have. A random forest algorithm will run quickly and accurately on a large dataset with hundreds of variables without variable deletion to provide us with a shortened list of the most important exposures in a foodborne outbreak.[9] It will also handle missing data, maintaining accuracy when a large portion of the data is missing. Random forests control for overfitting, allowing us to run as many trees as needed to produce relevant results.

While random forests is a popular tool in biomedical studies, it has not been used in foodborne epidemiological studies. With *Salmonella* being an ongoing public health burden and a major cause of outbreaks in the United States and South Carolina, prompt and effective detection of the source through outbreak investigations are necessary to remove contaminated food from the market, prevent further illnesses, and focus prevention strategies.[11] The primary aim of this study is to create a random forests algorithm using the hypothesis-generating questionnaire and laboratory data obtained from the South Carolina Department of Health and Environmental Control (DHEC) associated with previously identified *Salmonella* outbreaks in the state to identify the exposure source. A secondary aim of this study is to use variable importance to limit the current list of 207 individual food and environmental exposures to the top 30 suspected exposures that may have led to the outbreak to aid in subsequent follow-up interviews.

4.3 METHODS

4.3a STUDY PARTICIPANTS

DHEC staff attempted to call all South Carolinians who went to see a medical physician for symptoms of foodborne illness, submitted a stool sample for *Salmonella* testing, and received a confirmation of illness were eligible to fill out the state-mandated hypothesis-generating questionnaire on foodborne illness (See appendix B). Not all eligible participants were administered the questionnaire as some were lost to follow-up and others did not fully complete the questionnaire. All completed questionnaires and serotypes were collected from eighteen counties (Pickens, Greenville, Spartanburg, Cherokee, Union, Richland, Lexington, Newberry, Fairfield, Chester, Lancaster, York, Horry, Williamsburg, Georgetown, Berkeley, Dorchester, Charleston). The data is collected by the county of residence of the case. The counties were selected for this project based upon: 1) disease burden (66% of all cases of *Salmonella*, Shiga toxin-producing Escherichia coli (STEC) and *Listeria*); 2) geographic diversity (Upstate, Midlands and Coastal areas); 3) population density (65% of the total state population); and, 4) presence of three metropolitan statistical areas (MSAs) listed in the "Top 100" nationally within the selected counties.

The following events had to have occurred for an illness to be ascertained and thus, for inclusion into the study. The ill person must have sought medical care and a specimen must be submitted to DHEC Bureau of Laboratories (BoL) for testing. For the purposes of this analysis, only *Salmonella* cases were considered. The laboratory test must identify the causative agent as *Salmonella* and report the illness to DHEC Division of Acute Disease Epidemiology (DADE). DADE must then administer the appropriate questionnaire and the person who was ill must complete it fully. Participants in this research were grouped by serotype as recorded by BoL as well as whether or not they were part of an outbreak as determined by DADE. Grouping of serotypes was done by comparing dates of illness onset, serotype pattern, and whether or not this is a confirmed case as determined by BoL.

Data were collected for this study from January 2008 to June 2015, with 4058 possible *Salmonella* cases identified for inclusion based on the above criteria.

4.3b ASSESSMENT OF OUTCOME (OUTBREAK DEFINITION)

The FDA defines a foodborne outbreak as an incident in which two or more persons experience a similar illness after ingestion of a common food, and epidemiologic analysis implicates the food as the source of the illness.[12] DHEC assigned outbreak status in this dataset based on their knowledge of sporadically occurring cases. They limited it by geographical location, serotype, date, and foods eaten, taking into account seasonality and a threshold of commonly occurring cases. DHEC defines an outbreak as two or more persons not living in the same household with the same enteric illness following a common exposure.[13] This definition was provided to us in the data collected from DHEC for the entire study period with 78 identified individuals as part of an outbreak. Based on the definition provided, three known DHEC outbreaks were broken out for analysis, May 2015, August 2012, and May 2012. The May 2015 outbreak was suspected to be caused by fast food chain chicken with 24 suspected cases. The August 2012 outbreak had 17 suspected cases was thought to be caused by boiled peanuts, contact with an amphibian, or a waterborne exposure. The May 2012 outbreak was caused by guacamole, chips, and sour cream with 9 cases.[4]

4.3c ASSESSMENT OF EXPOSURE

The questionnaire used in this study is the hypothesis-generating questionnaire used when a positive laboratory test is reported to DADE. The questionnaire has 16 sections with 242 questions about food and environmental exposures and locations where exposure occurred. Each response was coded as yes, no, unknown, or refused to answer to whether they had eaten that food within the past 7 days (See Appendix B). We restricted analysis to 207 individual exposures that were either a food or an environmental exposure.

4.3d OTHER COVARIATES

Other variables used in analysis included demographic groups broken down as urban or rural environment, age, gender, Hispanic, and race. Urban or rural was categorized by using the county of residence from the questionnaire and then the rural definition based on the Office of Management and Budget (OMB) metropolitan statistical areas. A metropolitan area includes one or more counties containing a core urban area of 50,000 or more people, together with any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core. OMB also defines micropolitan statistical areas using the same method but centered on urban areas with at least 10,000 but no more than 50,000 people.[14]

Gender was classified as 0 for female and 1 for male for ease of analysis. Age was categorized as an integer with null representing all missing data. Race was made into a numeric variable and put into categories of white, African American, other, and unknown/refused. Hispanic was asked as whether or not the person being interviewed identified as Hispanic with the answers being categorized as 0 for no, 1 for yes, 99 and missing as unknown.

4.3e ANALYSIS

Using SAS 9.4 software, we cleaned the original DHEC dataset. We changed all numeric food and non-food exposures to have 3 levels, yes, no, and missing. A refusal or non-response to a question was considered missing. SAS software was also used to create individual outbreak files to be used in later analyses in R.[15] R version 3.1.3 was used to perform a random forests model with the aid of the randomForest package version 4.6-12.[16]

Random forests is a package contained in R that grows many classification trees. The core building block of random forests is a CART (classification and regression tree) inspired decision tree. The CART algorithm starts by drawing a random sample from the main dataset and building a decision tree based on this sample. Then, it repeats the process a second time, picking another random sample and growing a second decision tree. The prediction from the second tree will typically be different (at least a little) than those of the first tree.[17] This process continues, generating more trees each built on a slightly different sample and generating at least slightly different predictions each time. Random forests builds upon CART by adding randomness into the actual tree growing and not just the sampling.[18] In the normal process in growing a decision tree is to conduct exhaustive searches across all possible predictors to find the best possible partition of data in each node of the tree. Random forests sometimes picks the best split at random to guarantee the dissimilarity in trees.[17] This is done by selecting a new random subset of predictors in each node of a tree. Predictions about the best variables for the model are made by averaging the predictions made by the trees. We developed a random forests model of 500 trees for three separate *Salmonella* outbreak events that occurred in South Carolina to assess importance of food and/or environmental exposures.

Importance measures the strength of a variable to be included in the final model. It is based on measuring the damage that would be done to our predictive models if we lost access to true values of a given variable.[17] To simulate losing access to a predictor, values are randomly scrambled in the data. That is, the value belonging to a specific row of data is moved to another row.[17] This is done one predictor at a time and the loss in accuracy is measured. Random forests scrambles the data for each predictor being tested in every tree in the forest, which removes the dependence on luck of the draw predictions. For example, if a predictor is scrambled 500 times in front of 500 trees, the results should be highly reliable.[17] While that measures accuracy, there is another measure, GINI. GINI is based on the actual role of a predictor and offers an alternative importance assessment based on the role the predictor plays in the data. It is a measure of how often a chosen predictor would be incorrectly classified if it was classified at random based on the subset of data chosen at each tree.[19] GINI then calculates each predictor importance as the sum over the number of splits across all trees, giving a fast variable importance that is often very consistent with the permutation importance measure. Consequently, importance is not equivalent to an effect measure of the exposure on the outcome.[20] In this study, we were only interested in narrowing the list of potential exposures to aid in follow-up questionnaires.

4.4 RESULTS

Our dataset contained 4058 *Salmonella* cases identified by DHEC from 2008-2015. Of those, DHEC identified 50 as being part of one of three outbreaks used in the analysis. Table 4.1 shows the study population characteristics for the three separate outbreaks studied here. It was shown that in the three outbreaks, most were in an urban setting, non-Hispanic, female, and the race of most cases was white. With age, the outbreak in May 2015 affected people older than 45 while the outbreaks in August and May 2012 affected mostly people under 45.

4.4a VARIABLE IMPORTANCE

Three known outbreaks were run through the random forests model and the top 30 exposures are shown in Figures 4.1, 4.2, and 4.3. Each separate outbreak model combined both food and environmental exposures into one and top exposures varied widely per outbreak. With the May 2015 outbreak, breaded chicken appeared on the top 30 list in both accuracy and GINI indices (Tables 4.2 and 4.3), which correlates with the actual cause of the outbreak- fried, breaded chicken. The outbreak in August 2012 results in peanuts in the GINI index, but not in the accuracy index (Tables 4.4 and 4.5). Both however do show consumption of peanut products, like peanut butter and pre-packaged peanut butter crackers. Contact with an amphibian does not appear at the top for either index and DHEC does not track waterborne exposures on the hypothesis-generating questionnaire. Guacamole and chips were shown in the list of top 30 exposures as a suspect for the May 2012 outbreak for both accuracy and GINI indices (Tables 4.6 and 4.7). Other things eaten with guacamole were also shown in the list, like Mexican cheese and shredded cheese.

4.5 DISCUSSION

The model used in this study was able to provide a shortened list of exposures that could lead to faster follow-up studies, thus reducing the number of illnesses that could occur. By knowing the causes of the outbreaks before running the model, we were able to test the efficiency of random forests with a high number of exposures. Traditionally, tree-based models are structurally accommodating of conditional causality in which an exposure high on a tree is related to the disease risk through exposures down the tree.[20] It was found that the outbreaks seemed to affect urban, non-Hispanic, white females with age being under 45 for the 2012 outbreaks while the 2015 outbreak affected mostly those over 45. With the May 2015 outbreak, DHEC suspected chicken from a fast food chain as the cause. The random forests model placed breaded chicken 7th (increasing accuracy) and 5th (decreasing GINI). The outbreak in August 2012 did not have a definitive cause identified by DHEC but boiled peanuts, contact with an amphibian, or a waterborne exposure were suspected. Since the hypothesis-generating questionnaire does not deal specifically with waterborne exposures in a yes/no/don't know question, the model was looking for the first two exposures. It found peanuts to be in the top 30, but amphibian contact was not. The May 2012 outbreak was caused by guacamole and sour cream as identified by DHEC's current model of calculating odds ratios (Table 4.8). Although the results for guacamole, sour cream, and chips are statistically significant for a 95% confidence interval shown in Table 4.8, the small sample size and wide confidence intervals preclude drawing any firm conclusions as to the source of the contamination. Again, the hypothesis-generating questionnaire does not specifically have a question regarding sour cream, so we were concerned with chips and guacamole. Chips were 3rd on the list for both increasing accuracy and decreasing GINI while guacamole was 17^{th} and 25^{th}, respectively. The random forests model has led to a shortened list of food and environmental exposures, thus making follow-up easier and more efficient. DHEC would be able to take the results from the shortened follow-up and use them to potentially find the cause of the outbreak faster and reduce the number of subsequent illnesses.

4.5a STRENGTHS AND LIMITATIONS

In this study, outbreaks were identified and people were surveyed without delay since the state mandates that probable *Salmonella* cases be reported within 3 business days. This helps to avoid possible information and selection bias. However, this study is still subject to bias. Since the laboratory test must identify the causative agent serotype and report the illness to public health authorities, the appropriate questionnaire may not be administered in a timely fashion. Delays in reporting could cause recall bias when the ill person is attempting to answer all exposure questions. If they cannot recall what is eaten in the 7 days prior to becoming ill, they may answer "I don't know" or answer as to what is normally eaten on an average day. This could over or underestimate the risk of certain foods eaten, leading to misclassification of foods eaten prior to illness and spurious associations could arise. It has been shown that foods eaten in a particular week are correlated with what is eaten in the past.[21] If the ill person ate something out of character, it is likely they would remember since it possibly led to their illness. Because of this, it is unlikely that these errors would have affected the overall shortened list of exposures per outbreak.

Also, random forests maintain their accuracy when a large proportion of the data are missing, something very vital in foodborne outbreak investigations.[9] Random forests can also handle the large number of variables input into the model without variable deletion. This is especially important in our model since any one of the 207 exposures could potentially be the cause of the outbreak. Another advantage of using random forests is that it was able to handle the small sample sizes input to the model. The outbreaks consisted of samples of 24 cases (May 2015), 17 cases (August 2012), and 9 cases (May 2012). The model ran efficiently even with the limited number of cases.

A limitation of this study is that random cannot exactly predict the true exposure that caused the outbreak. This could partially be due to randomness in the model, which could result in a random variable having a stronger value of importance than the true outbreak exposure. Our results did put the actual outbreak cause toward the top of the exposure list, but it was by no means the top answer. While we may not be able to predict exactly, the random forests model can focus in on causal inference by using measurements of importance. In this study, we were able to create a more concise list of outbreak exposures, taking it from 207 to 30 within a matter of minutes. This will allow public health epidemiologists to have a more focused follow-up investigation in hopes of reducing the number of illnesses associated with the contaminated exposure. Random forests has advantages over current methods in detecting exposures in foodborne outbreaks, like regression models. This due to random forest's ability to be flexible, model interactions, and the way it actually handles the missing data throughout the model.

4.6 CONCLUSION

In this study, we used random forests to model food and environmental exposures obtained from a hypothesis-generating questionnaire to create a shortened list of causes of outbreaks. We used random forests because it can learn non-linear relationships and interactions from data, which was useful in this study due to the large number of exposure variables.[10] Logistic regression is commonly used in epidemiological studies as it can provide relative risk and odds ratios, but it can be limited when analyzing datasets like the one in this study because of its high number of exposures and the multiple interactions they can have. It would have been nearly impossible to add all relevant interactions to a logistic regression model considering there were 207 relevant exposures. Random forests is a popular method used in biomedical studies, but its use in epidemiological studies is minimal.[7,8,22,23] We applied the random forests method to analyze the complex relationships in food outbreak data, handle a high volume of exposures, and deal with missing data.

We used a random forests algorithm to find exposures for cases in three *Salmonella* outbreaks that occurred in South Carolina in the past 5 years. The random forests algorithm generated lists of the top 30 suspected exposures out of 207 individual food and environmental exposures that contained the foods associated with each of the 3 outbreaks that were analyzed. Random forests may aid in investigations of foodborne outbreaks and aid in quicker identification of their causes.

4.7 WORKS CITED

1. Centers for Disease Control and Prevention (CDC). *Salmonella.* https://www.cdc.gov/salmonella/. Published 2016. Accessed 18 Oct 2016.

2. National Institute of Diabetes and Digestive and Kidney Diseases. Foodborne Illnesses. http://digestive.niddk.nih.gov/ddiseases/pubs/bacteria/.Published 2012. Accessed 18 Oct 2016.
3. Santos RL, Raffatellu M, Bevins CL, Adams LG, Tükel C, Tsolis RM, Bäumler AJ. Life in the inflamed intestine, *Salmonella* style. Trends Microbiol. 2009:17, 498–506.
4. South Carolina Department of Health and Environmental Control. DHEC Hypothesis Generating Questionnaire. http://www.scdhec.gov/library/D-0928.pdf. Published2011. Accessed 31 Oct 2016.
5. Centers for Disease Control and Prevention. Estimates of Foodborne Illness in the United States. http://www.cdc.gov/foodborneburden/estimates-overview.html. Published 2016. Access 1 Nov 2016.
6. Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiology.* 2012 Jan; 23(1): 159–164.
7. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics.* 2006; 7:3.
8. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics.* 2007; 23(19): 2507-2517.
9. Breiman L. Random forests. Machine Learning 2001; 45: 5–32.
10. Gromping U. Variable importance assessment in regression: linear regression versus random forest. *American Statistician.* 2009; 63: 308–319.
11. Centers for Disease Control and Prevention. Burden of Foodborne Illness: Questions and Answers. https://www.cdc.gov/foodborneburden/questions-and-answers.html. Published 2016. Accessed 5 Feb 2017.
12. U.S. Food and Drug Administration. What You Should Know About Government Response to Foodborne Illness Outbreaks. http://www.fda.gov/Food/ResourcesForYou/Consumers/ucm180323.htm. Published 2014. Accessed 20 October 2014.
13. Council to Improve Foodborne Outbreak Response (CIFOR). Guidelines for Foodborne Disease Outbreak Response. http://www.cifor.us/CIFORGuidelinesProjectMore.cfm. Published 2014. Accessed 3 Apr 2017.
14. U.S. Department of Agriculture Economic Research Service. South Carolina - Rural Definitions: State Level Maps. http://www.ers.usda.gov/datafiles/Rural_Definitions/StateLevel_Maps/SC.pdf. Published 2015. Accessed 15 Nov 2015.
15. SAS Institute Inc 2013. SAS/ACCESS® 9.4 Interface to ADABAS: Reference. Cary, NC: SAS Institute Inc.
16. RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.
17. Salford Systems. Random Forests for Beginners. http://info.salford-systems.com/an-introduction-to-random-forests-for-beginners. Published 2014. Accessed 19 Feb 2017
18. UC Berkeley. Random Forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. Published 2001. Accessed 20 Sep 2014.

19. Strickland J. "9. Decision Tree Learning." Data Analytics Using Open-Source Tools. N.p.: Lulu.com, 2016. 201.
20. Gu W, Vieira AR, Hoekstra RM, Griffin PM, and Cole D. Use of random forest to estimate population attributable fractions from a case-control study of *Salmonella enterica* serotype Enteritidis infections. *Epidemiology and infection.* 2015; 143(13): 2786-2794.
21. Rimm ED, Giovannucci EL, Stampfer MJ, Colitz GC, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. *Am J Epidemiol* . 1992;135: 1114-1126.
22. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*. 2007; 8: 25.
23. Boulesteix AL, Bender A, Bermejo JL, Strobl C. Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics.* 2012; 13 (3): 292-304.

4.8 TABLES

TABLE 4.1 STUDY POPULATION CHARACTERISTICS OF THE 3 OUTBREAKS

Value	Total	Yes		Age		Gender		Hispanic		Race		Urban/Rural	
	N	N (YES)	%	N (<45)	%	N (Male)	%	N (YES)	%	N (White)	%	N (Urban)	%
Outbreak May 2015: Fast Food Chain Chicken	24	19	79.17%	5	20.83%	7	29.17%	0	0.00%	14	58.33%	16	66.67%
Outbreak August 2012: Boiled Peanuts, Amphibian Exposure	17	16	94.12%	9	52.94%	8	47.06%	2	11.76%	12	70.59%	16	94.12%
Outbreak May 2012: Guacamole, Sour Cream, Chips	9	7	77.78%	5	55.56%	1	11.11%	0	0.00%	7	77.78%	7	77.78%

TABLE 4.2 VARIABLE IMPORTANCE FOR THE MAY 2015 OUTBREAK WITH DECREASING ACCURACY

VARIABLE	MEANDECREASEACCURACY
UNKNOWN MILK	3.31901015
APPLE JUICE	2.88826280
HEAD OF ICEBERG	2.82549247
SLICED CHEESE	2.78907083
GRANOLA	2.35505601
LOOSE GREENS	2.11709234
BREADED CHICKEN	2.02105463
CONTACT WITH A DOG	1.89519850
CONTACT WITH A CAT	1.85562869
WHOLE TURKEY	1.83890609
CRAB	1.72420382
RUNNY EGGS	1.69476283
VISITED A PET STORE	1.68788787
ICEBERG	1.62665271
LETTUCE ON A SANDWICH	1.59227494
GRAPE	1.56791732
GROUND BEEF DISH AWAY FROM HOME	1.54075934
MUSHROOM	1.25176716
STEAK	1.14284465
PEPPERONI	1.00100150
PINEAPPLE	1.00100150
GUACAMOLE	1.00100150
PRE-PACKAGED GREENS	1.00100150
CHOCOLATE	0.85401075
PEANUT	0.82612522
FRESH STEAK AT HOME	0.49976734
CASHEW	0.44487760
ALMOND	0.42527210
PRE-PACKAGED PEANUT BUTTER CRACKERS	0.40704706
CREAMY PEANUT BUTTER	0.37224525

TABLE 4.3 VARIABLE IMPORTANCE FOR THE MAY 2015 OUTBREAK WITH DECREASING GINI

VARIABLE	MEANDECREASEGINI
UNKNOWN MILK	0.30799550
SLICED CHEESE	0.29947210
WHOLE TURKEY	0.20302220
GROUND BEEF DISH AWAY FROM HOME	0.20200480
BREADED CHICKEN	0.18450130
CHIPS	0.17903680
PEANUT	0.16840120
CONTACT WITH LIVE POULTRY	0.16564080
OTHER JUICE	0.16088970
CRUNCHY PEANUT BUTTER	0.15276030
EGGS	0.15010410
FROZEN STEAK AT HOME	0.13766300
HEAD OF ICEBERG	0.13740950
OTHER DELI MEAT	0.13426570
LOOSE GREENS	0.13057630
FRUIT ROLL	0.12442180
PREFORMED GROUND BEEF PATTY	0.11825250
MILK	0.11774120
FROZEN SNACKS	0.11535250
CANTALOUPE	0.11304300
PARMESEAN	0.10483550
GREENS	0.10466030
UNKNOW CONTACT WITH PET FOOD	0.10172400
WATERMELON	0.10000960
PRE-PACKAGED PEANUT BUTTER CRACKERS	0.08761465
HOTDOGS	0.08653445
FRESH GROUND BEEF PATTY	0.08531437
BACON	0.08382177
FROZEN VEGETABLES	0.08294094
PRE-PACKAGED CRACKERS	0.08171168

TABLE 4.4 VARIABLE IMPORTANCE FOR THE AUGUST 2012 OUTBREAK WITH DECREASING ACCURACY

VARIABLE	MEANDECREASEACCURACY
PRE-PACKAGED CRACKERS	1.69066636
APPLE	1.40275209
ANY PEANUT BUTTER	1.28845246
CEREAL	1.13907104
ICEBERG	1.13828629
PARMESEAN	1.00100150
BLUEBERRY	1.00100150
CHOCOLATE	1.00100150
CONTACT WITH A CAT	1.00100150
OTHER YOGURT	0.97897458
GRAPE	0.86061028
HOT CEREAL	0.75979081
PEPPERONI	0.57028712
FROZEN VEGETABLES	0.32016518
CONTACT WITH A DOG	0.27832251
FRESH TOMATO	0.25817970
HEAD OF ICEBERG	0.20373983
EGGS	0.14554185
GREENS	0.12803898
FRUIT ROLL	0.02461086
GROUND CHICKEN	0.00000000
CHICKEN KIEV	0.00000000
OTHER FROZEN CHICKEN	0.00000000
DUCK	0.00000000
WHOLE TURKEY	0.00000000
GROUND TURKEY	0.00000000
FROZEN STEAK AT HOME	0.00000000
PINK STEAK AT HOME	0.00000000
PREFORMED GROUND BEEF PATTY	0.00000000
PREFORMED PINK PATTY	0.00000000

TABLE 4.5 VARIABLE IMPORTANCE FOR THE AUGUST 2012 OUTBREAK WITH DECREASING GINI

VARIABLE	MEANDECREASEGINI
UNKNOWN CONTACT WITH PET FOOD	0.56119940
MINI CARROT	0.12953560
CONTACT WITH A DOG	0.08589894
OTHER PORK	0.07478352
ANY PEANUT BUTTER	0.07219670
SLICED CHEESE	0.07212952
HOT CEREAL	0.05646376
CREAMY PEANUT BUTTER	0.04738312
STEAK	0.04702633
ICEBERG	0.04653310
FRESH TOMATO	0.04088795
OTHER FROZEN ITEM	0.03951212
FRESH FISH	0.03563516
GRAPE	0.03287119
OTHER GREENS	0.03058990
HEAD OF ICEBERG	0.02968739
PEANUT	0.02711970
FROZEN VEGETABLES	0.02526661
GROUND BEEF DISH AT HOME	0.02481172
CEREAL	0.02407370
HOT DOGS	0.02372028
RED ROUND TOMATO	0.02310299
SAUSAGE	0.02212047
EGGS	0.02124196
SHREDDED CHEESE	0.02119925
CHIPS	0.02064225
PRE-PACKAGED DELI MEAT	0.01905983
PRE-PACKAGED PET FOOD	0.01871480
BROCCOLI	0.01856873
BOLOGNA	0.01575482

TABLE 4.6 VARIABLE IMPORTANCE FOR THE MAY 2012 OUTBREAK WITH DECREASING ACCURACY

VARIABLE	MEANDECREASEACCURACY
SUNFLOWER	3.79923640
SHREDDED CHEESE	3.47352600
CHIPS	3.46088450
MILK	2.22660440
BREADED CHICKEN	2.07360060
PEPPERONI	1.97158040
WHOLE CHICKEN	1.66911010
MEXICAN CHEESE	1.41705050
SLICED CHEESE	1.41705050
FRESH GROUND BEEF PATTY	1.32225570
FRESH FISH	1.00100150
CREAMY PEANUT BUTTER	1.00100150
CONTACT WITH PET TREATS	1.00100150
CANTALOUPE	1.00100150
SAUSAGE	1.00100150
CEREAL	0.46862420
GUACAMOLE	0.42647900
STEAK	0.24254990
HOT CEREAL	0.00000000
OJ	0.00000000
CONTACT WITH A CAT	0.00000000
BLOCK CHEESE	0.00000000
OTHER GREENS	0.00000000
HEAD OF ICEBERG	0.00000000
SCALLION	0.00000000
OTHER DELI MEAT	0.00000000
CASHEW	0.00000000
OTHER PORK	0.00000000
GROUND BEEF DISH AWAY FROM HOME	0.00000000
ROMAINE	0.00000000

TABLE 4.7 VARIABLE IMPORTANCE FOR THE MAY 2012 OUTBREAK WITH DECREASING GINI

VARIABLE	MEANDECREASEGINI
SHREDDED CHEESE	0.27039524
SUNFLOWER	0.24130794
CHIPS	0.22911111
MILK	0.11728571
BREADED CHICKEN	0.10652857
HOT CEREAL	0.09442063
OJ	0.09226825
CONTACT WITH A CAT	0.08980952
BLOCK CHEESE	0.08068889
PRE-PACKAGED PEANUT BUTTER CRACKERS	0.07867143
FRESH GROUND BEEF PATTY	0.07729841
OTHER GREENS	0.07245238
STRAWBERRY	0.06711746
OTHER YOGURT	0.06465238
WHOLE CHICKEN	0.06390317
HEAD OF ICEBERG	0.05829206
PEANUT	0.05811587
SCALLION	0.05623810
EGGS	0.05587937
CEREAL	0.05082222
FRESH FISH	0.04951587
OTHER DELI MEAT	0.04596032
CASHEW	0.03967460
MEXICAN CHEESE	0.03905714
GUACAMOLE	0.03662857
OTHER PORK	0.03652222
PEPPERONI	0.03460952
BACON	0.03358095
BELL PEPPER	0.03330794
GROUND BEEF DISH AWAY FROM HOME	0.03216667

TABLE 4.8 CAUSES OF THE MAY 2012 OUTBREAK AS IDENTIFIED BY DHEC

VARIABLE	ODDS RATIO	CONFIDENCE INTERVAL
SOUR CREAM	5.33	0.9 - 31.9
SUNFLOWER	8.00	1.2 - 51.5
CHIPS	8.00	1.2 - 51.5

4.9 FIGURES

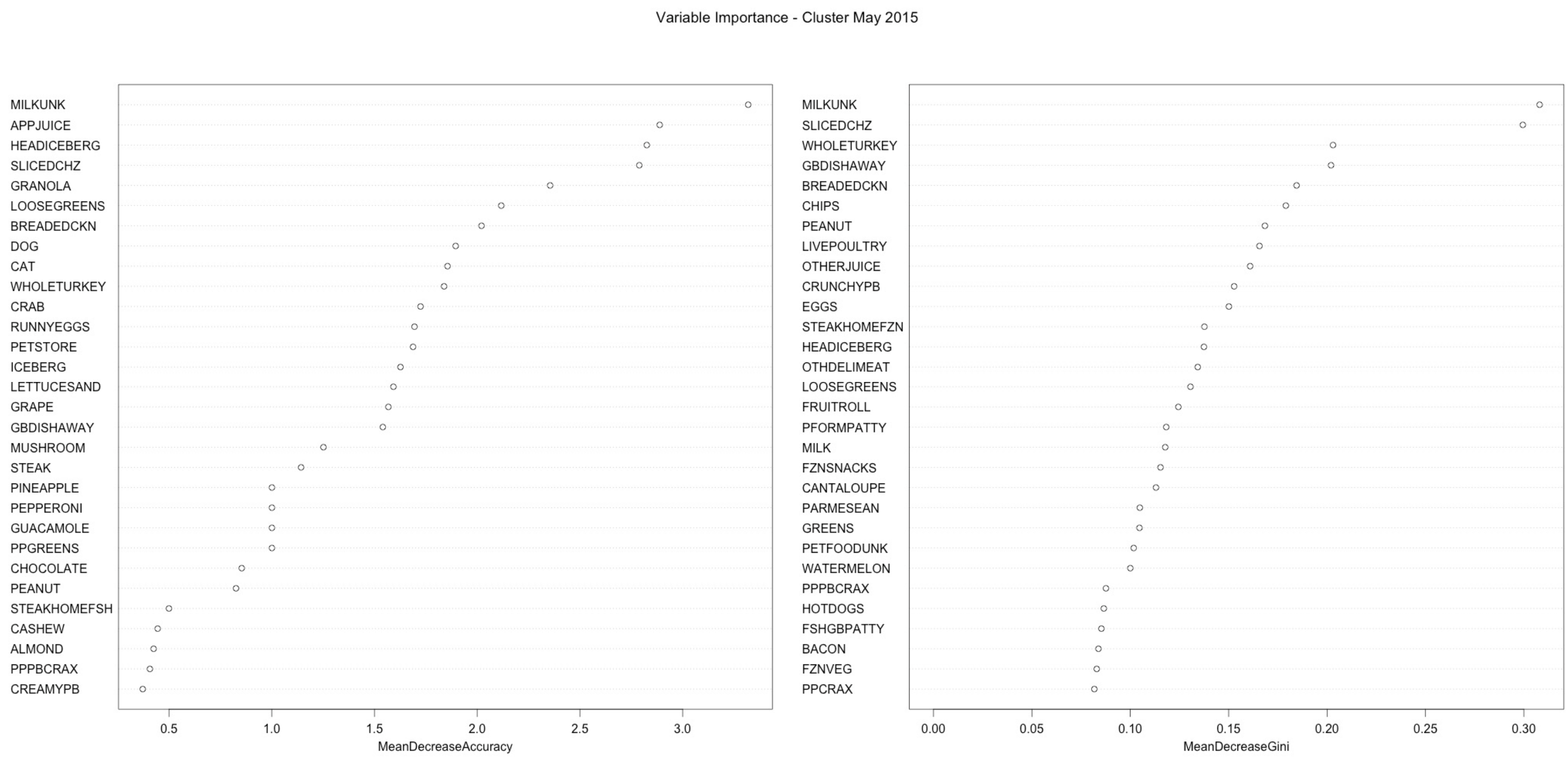

Figure 4.1 Variable Importance for the May 2015 Outbreak - This should be interpreted in terms of decreasing importance. The higher the variable is on the list, the more important it is to the model.

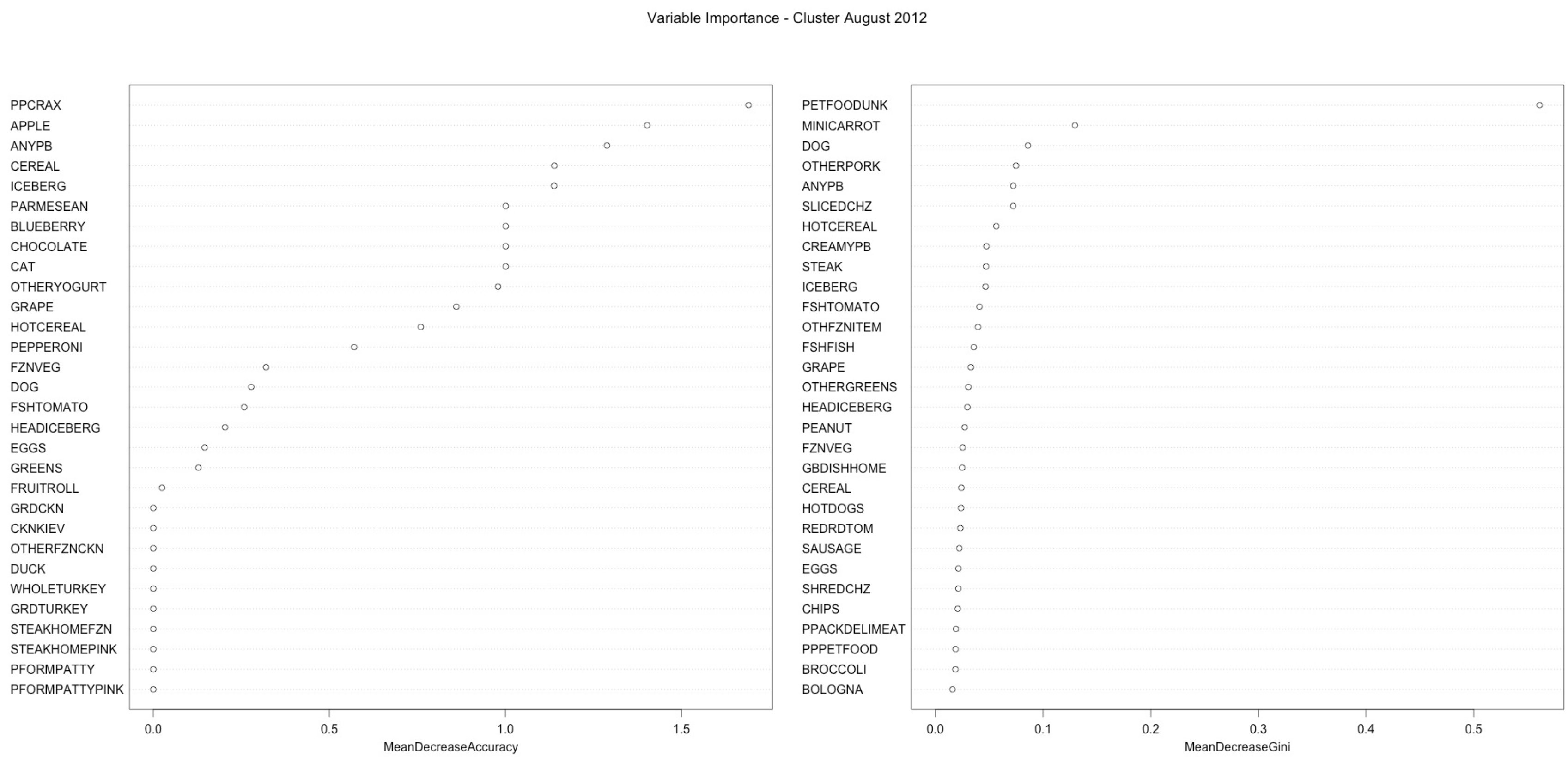

Figure 4.2 Variable Importance for the August 2012 Outbreak - This should be interpreted in terms of decreasing importance. The higher the variable is on the list, the more important it is to the model.

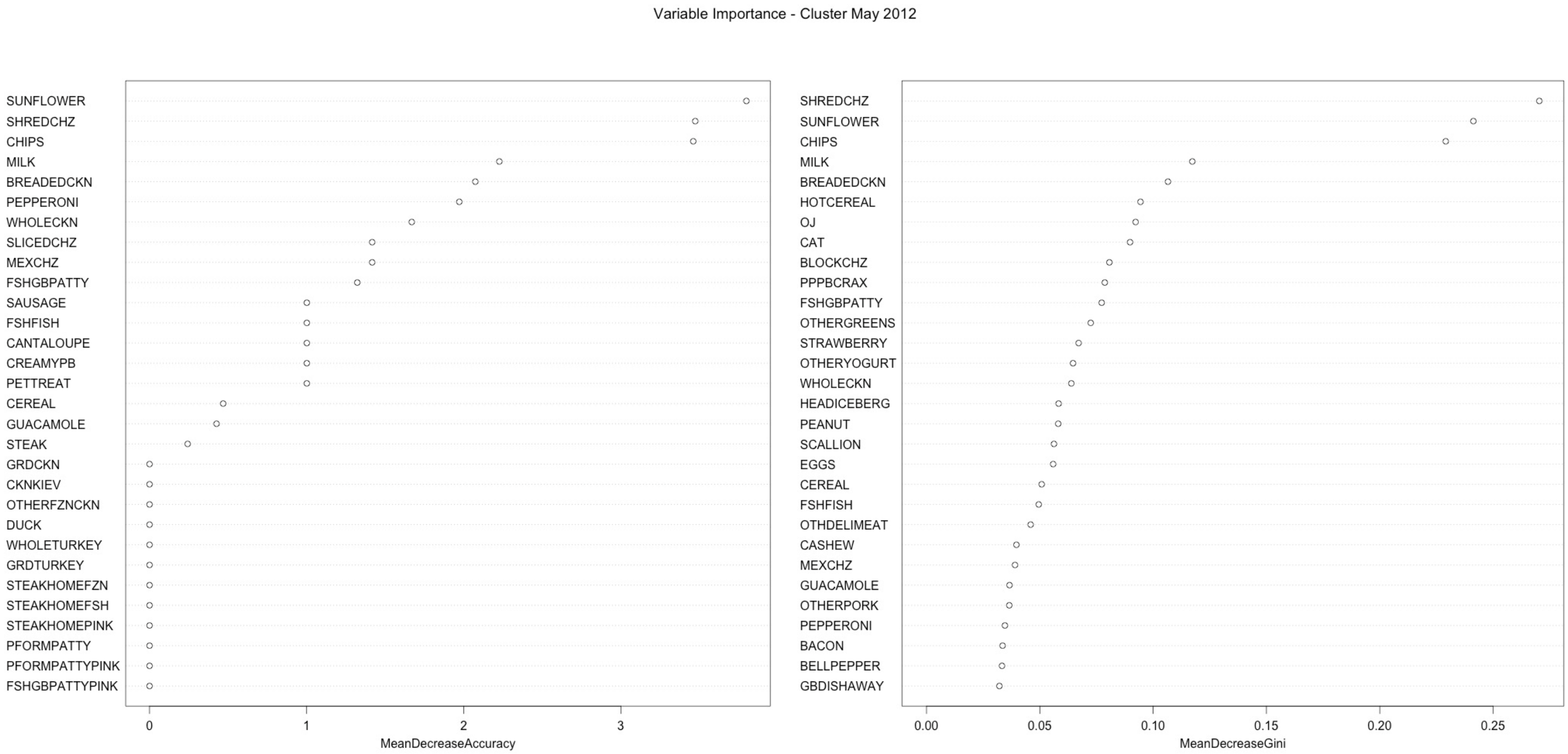

Figure 4.3 Variable Importance for the May 2012 Outbreak - This should be interpreted in terms of decreasing importance. The higher the variable is on the list, the more important it is to the model.

CHAPTER 5

EVALUATION OF FOOD CAUSES OF *SALMONELLA* OUTBREAKS IN SOUTH CAROLINA[2]

[2] Alianell, A.T., Merchant, A., McLain, A., Brenner, E., and D. Giurgiutiu. To be submitted to *American Journal of Epidemiology*.

5.1 ABSTRACT

Foodborne illness is a serious public health threat, with causes varying greatly throughout the state of South Carolina. Food history is an important starting point in outbreak investigations. Because time is a major interest in investigations, we evaluated methods that may aid in quick identification of foods that may be associated with outbreaks. We did this by examining previous cases of salmonellosis using 2 different definitions (state and FDA) of what constitutes an outbreak. Using data provided by the South Carolina Department of Health and Environmental Control (DHEC), there were 78 DHEC defined outbreak associated cases and 2565 potential outbreak associated cases using the FDA definition. After conducting various models, the DHEC outbreak definition identified meat as 2.78 times as likely to be associated with an outbreak, dairy as 0.52 times as likely, and greens as 2.5 times as likely, with urban/rural, and Hispanic being demographic indicators of significance. The FDA outbreak definition only identified dairy as 1.3 times as likely to be associated with an outbreak of foodborne illness in South Carolina. This investigation showed that there are many differences in both the number of potential cases identified and the outbreak associated foods provided by the two definitions. Due to the high number of differences in the definitions, further investigation will be needed to address the challenges of a rural southern state with high disease burden in locations with the highest number of cases.

5.2 INTRODUCTION

Foodborne illness is a serious public health threat. The Centers for Disease Control and Prevention (CDC) estimates that 76 million foodborne illnesses, including 325,000 hospitalizations and 5,000 deaths, occur in the United States each year.[1,2] This roughly equates to 1 in 6 Americans getting sick every year.[2,3] Foodborne illness does not discriminate; anyone can get a foodborne illness. However, some people are more likely to develop foodborne illnesses than others, including infants and children, pregnant women, older adults, and people with weak immune systems.[4] Many different disease-causing microbes (or pathogens) can contaminate foods or beverages, leading to various different foodborne infections.[3] Most diseases are infections caused by a variety of bacteria, viruses, and parasites that can be foodborne. Since so many microbes can cause foodborne illness, there can be many different symptoms, but the most common are nausea, vomiting, abdominal cramps, and diarrhea.[3,5] Most healthy individuals will recover without treatment in about 4 to 7 days without the need for hospitalization.[6]

One of the most common causes of foodborne illness is *Salmonella*. *Salmonella* is a Gram-negative facultative intracellular pathogen that causes a spectrum of clinical diseases depending on the serotype of the infecting bacteria and the susceptibility of the host.[7,8] Infections fall into three categories, (1) gastroenteritis, (2) systemic infection of an otherwise healthy host, or typhoid, and (3) infection of an immunocompromised host. In terms of this review, focus will be on the first of three types, gastroenteritis, or non-typhoidal *Salmonella*.

Large foodborne outbreaks, including those caused by *Salmonella*, associated with the ingestion of contaminated foods, like tomatoes, produce, and peanut butter are becoming a more and more common occurrence.[7,9-12] Approximately 50% of all foodborne infections are caused by bacteria, and of those, about 30-50% can be attributed to *Salmonella* and its variates.[1,13] What is important to note is that these intestinal infections can be initiated by any of the approximately 2,000 different serotypes of *Salmonella* that infect both human and animal reservoirs.[7]

The most common human isolates of non-typhoidal *Salmonella* (NTS) are *Salmonella* Enterica serotypes Typhimurium (*S.* Typhimurium) and Enteritidis (*S.* Enteritidis) as well as *S.* Newport *and S.* Heidelberg.[14] Annually, there are around 40,000 cases of NTS reported each year, which underestimates the actual problem due to the ill person not visiting a physician or the fact that no specimen is obtained for laboratory testing.[15,16] It is estimated that of these cases, 582 deaths occur each year, making this pathogen the leading cause of foodborne infections with lethal outcomes in the United States.[14] The incidence of foodborne human infections caused by *S.* Enteritidis and by multi-drug-resistant strains of *S.* Typhimurium increased substantially during the second half of the 20th century moving into the first part of the 21st century in the US with similar trends being reported from Europe.[17-21]

With foodborne outbreaks of *Salmonella* constituting a major, ongoing public health burden in the United States and South Carolina, prompt and effective detection of the source through outbreak investigations are necessary to remove contaminated food from the market, prevent further illnesses, and focus prevention strategies on critical contamination points along the "farm-to-fork" continuum. In South Carolina, an ill person must seek medical care and a specimen must be submitted to DHEC Bureau of Laboratories (BoL) for testing for surveillance of foodborne illnesses to begin. The laboratory test must identify the causative agent and report the illness to public health authorities, in this case DHEC Division of Acute Disease Epidemiology (DADE). DADE must then administer the appropriate questionnaire and the person who was ill must complete it fully. The aim of the current study is to evaluate the foods associated with *Salmonella* outbreaks in South Carolinians using two standard definitions of an outbreak; the first definition is the one used by DHEC to identify outbreaks and the second is used by the FDA as a general definition of a foodborne outbreak. We aim to look at commonalities and differences in the two definitions using model selection using food data obtained using questionnaires from residents of South Carolina. This information will address the challenges of a rural southern state with high disease burden in locations with the highest number of cases, something that has not previously been shown.

5.3 METHODS

5.3a STUDY PARTICIPANTS

Participants were included if they sought medical care in one of 18 counties (Pickens, Greenville, Spartanburg, Cherokee, Union, Richland, Lexington, Newberry, Fairfield, Chester, Lancaster, York, Horry, Williamsburg, Georgetown, Berkeley, Dorchester, Charleston), that were selected for this sentinel project. They were selected based upon: 1) disease burden (they account for 66% of all cases of *Salmonella*, Shiga toxin-producing Escherichia coli (STEC) and *Listeria*), 2) geographic diversity (Upstate, Midlands and Coastal areas), 3) population density (65% of the total state population), and, 4) presence of three metropolitan statistical areas (MSAs) listed in the "Top 100" nationally.

The following events must occur for an illness to be ascertained and thus inclusion in the study. The ill person must seek medical care and a specimen must be submitted to DHEC Bureau of Laboratories (BoL) for testing. The laboratory test must identify the causative agent and report the illness to DHEC Division of Acute Disease Epidemiology (DADE). DADE must then administer the appropriate questionnaire and the person who was ill must complete it fully. They will be grouped by serotype as recorded by BoL as well as whether or not they were part of an outbreak as determined by DADE. Grouping of serotypes was done by comparing dates of illness onset, serotype pattern, and whether or not this is a confirmed case as determined by BoL.

Data were collected for this study from January 2008 to June 2015, with 4058 possible *Salmonella* cases identified for inclusion based on the above criteria.

5.3b ASSESSMENT OF OUTCOME (OUTBREAK AND FOOD GROUP DEFINITIONS)

The FDA defines an outbreak as two or more cases of foodborne illness caused by the same organism that occur within a limited period of time and are associated with either the same food or same food service operation.[22] This study interpreted this as two or more cases of the same organism that occurred within 30 days and used this to find the food commonalities. Since DHEC has more knowledge of sporadically occurring cases, they are more stringent with their outbreak definition and limit it by serotype, date, and foods eaten. This definition was provided to us in the data collected from DHEC. Currently, this is being looked at manually at DHEC and this study is doing it programmatically using SAS 9.4 for all analyses.

5.3c ASSESSMENT OF EXPOSURE

Currently, it is mandated by the state of South Carolina that cases of *Salmonella* need to be reported to DHEC within 3 business days. While many more pathogens cause gastrointestinal illness, the main focus here will be *Salmonella* as it is one of the top causes of illness in the US. This requires daily monitoring of laboratory and provider disease reports to identify cases followed by immediate interview of identified cases. The interview includes a standard core questionnaire assessing food history for initial screening and hypothesis generation for all diagnosed cases of infection with *Salmonella*.

Once a possible case is identified, a member of the DHEC staff would attempt to contact the case and conduct the interview in person by reading the questions to the possible case and marking the correct answer. The questionnaire consisted of a list of commonly eaten foods and the respondent was required to answer yes, no, unknown, or refused to answer to whether they had eaten that food within the past 7 days. An example of the questionnaire is included as Appendix B. Interviews were done without waiting for the serotype results. Conducting real-time review of subtyping results in conjunction with the interviews made it possible to see cluster-associated cases are evaluated together.

5.3d CATEGORIZATION OF FOODS

Foods eaten were grouped into specific categories for ease of analysis. The groups are poultry, meat, pork, seafood, egg, dairy, fruit, vegetables, greens, snacks, nuts, grains, raw foods, frozen foods, and prepackaged foods that are not frozen. The raw, frozen, and prepackaged categories may contain foods that are already in the main categories. They were split for further analysis.

The food grouping categories are listed below.

- Poultry: Whole Chicken, Ground Chicken, Breaded Chicken, Chicken Kiev, Other Frozen Chicken, Duck, Whole Turkey, Ground Turkey

- Meat: Steak, Frozen Steak Eaten at Home, Fresh Steak Eaten at Home, Pink Steak Eaten at Home, Preformed Patties, Pink Preformed Patties, Fresh Ground Beef Patties, Pink Fresh Ground Beef Patties, Any Ground Beef Dish Eaten at Home, Any Ground Beef Dish Eaten Away from Home, Lamb, Italian Meats, Bologna, Jerky, Pre-Packaged Deli Meat, Other Deli Meat, Any Other Meat
- Pork: Ground Pork, Other Pork, Bacon, Sausage, Hot Dogs, Pepperoni
- Seafood: Fresh Fish, Dried Fish, Shrimp, Crab, Oysters, Clams, Sushi, Frozen Fish, Any Other Seafood
- Egg: Eggs, Runny Eggs, Raw Eggs
- Dairy: Milk, Unknown Type of Milk, Raw Milk, Ice Cream, Frozen Yogurt, Yogurt Drinks, Other Yogurt, Shredded Cheese, Sliced Cheese, Block Cheese, String Cheese, Cottage Cheese, Cheese Curds, Feta, Blue Cheese, Parmesan, Raw Cheese, Mexican Cheese, Homemade Mexican Cheese, Gourmet Cheese, Dry Buttermilk, Flavored Milk Powder, Other Powdered Milk, Any Other Dairy
- Fruit: Apple, Grape, Pear, Peach, Nectarine, Apricot, Plum, Orange, Grapefruit, Tangerine, Lemon/Lime, Strawberry, Raspberry, Blueberry, Blackberry, Cherry, Any Other Berry, Cantaloupe, Honeydew, Watermelon, Precut Melon, Any Other Melon, Pineapple, Mango, Coconut, Any Other Tropical Fruit, Frozen Berries, Any Other Frozen Fruit, Raisins, Any Other Dried Fruit, Apple Juice, Orange Juice, Any Other Juice, Frozen from Concentrate Juice, Raw Juice

- Vegetables: Fresh Tomato, Red Round Tomato, Roma Tomato, Cherry Tomato, Grape Tomato, Tomatoes on the Vine, Any Other Tomato, Unknown Tomato, Tomato on a Sandwich, Salsa, Guacamole, Alfalfa, Sprouts, Cucumber, Bell Pepper, Hot Pepper, Celery, Mini Carrots, Other Carrots, Root Vegetables, Peas, Broccoli, Cauliflower, Onion, Scallion, Mushroom, Frozen Vegetables
- Greens: Greens, Pre-packaged Greens, Loose Greens, Lettuce on a Sandwich, Iceberg, Pre-packaged Iceberg, Head of Iceberg, Unknown Iceberg, Romaine, Pre-packaged Romaine, Loose Romaine, Spinach, Pre-packaged Spinach, Loose Spinach, Cabbage, Other Greens, Basil, Cilantro, Other Herbs
- Snacks: Frozen Pot Pie, Frozen Pizza, Frozen Mexican Food, Frozen Snacks, Frozen Breakfast, Frozen Vegetarian, Frozen Bagged Meal, Frozen Dinner, Other Frozen Item, Other Unknown Frozen Item, Pre-packaged Peanut Butter Crackers, Fruit Roll-up, Chips, Pre-packaged Crackers, Chocolate, Trail Mix, Powdered Nutrition Supplement
- Nuts: Any Peanut Butter, Creamy Peanut Butter, Crunchy Peanut Butter, Unknown Peanut Butter, Foods Containing Peanut Butter, Unknown Foods Containing Peanut Butter, Nut Butter, Peanuts, Almonds, Walnuts, Cashews, Pistachios, Hazelnut, Other Nuts, Sunflower Seeds, Sesame Seeds, Tahini, Hummus
- Grains: Granola, Cereal, Hot Cereal

5.3e OTHER COVARIATES

Other variables used in analysis included demographic groups broken down as urban or rural environment, age, gender, Hispanic, and race. Urban or rural was categorized by using the county of residence from the questionnaire and then the rural definition based on the Office of Management and Budget (OMB) metropolitan statistical areas. A metro area includes one or more counties containing a core urban area of 50,000 or more people, together with any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core. OMB also defines micropolitan statistical areas using the same method but centered on urban areas with at least 10,000 but no more than 50,000 people.[24]

Gender was classified as 0 for female and 1 for male for ease of analysis. Age was categorized as an integer with the option of being missing if the participant refused. Race was made into a numeric variable and put into categories of white, African American, other, and unknown/refused. Hispanic was asked as whether or not the person being interviewed identified as Hispanic with the answers being categorized as 0 for no, 1 for yes, 99 and missing as unknown.

If questions in the questionnaire specifically asked for frozen foods, raw foods, or pre-packaged foods, they were placed into groups as well. These groups overlap with the food groups themselves and were created as such:

- Frozen: Chicken Kiev, Other Frozen Chicken, Frozen Fish, Ice Cream, Frozen Yogurt, Frozen Berry, Other Frozen Fruit, Frozen Concentrated Juice, Frozen Pot Pie, Frozen Pizza, Frozen Mexican Food, Frozen Snacks, Frozen Breakfast, Frozen Vegetarian Meal, Frozen Bagged Meal, Frozen Dinner, Other Frozen Item, Frozen Steak at Home

- Raw: Sushi, Raw Egg, Raw Cheese, Raw Milk, Raw Juice, Root Vegetables, Peas, Onions, Scallions, Pink Steak at Home, Pink Pre-Formed Patties, Pink Fresh Ground Beef Patties, Runny Eggs, Raw Dairy
- Pre-Packaged: Jerky, Pre-Packaged Deli Meat, Shredded Cheese, Precut Melon, Pre-Packaged Greens, Pre-Packaged Iceberg, Pre-Packaged Romaine, Pre-Packaged Spinach, Pre-Packaged Peanut Butter Crackers, Granola, Trail Mix, Fruit Roll-Ups, Chips, Pre-Packaged Crackers, Chocolate, Cereal, Hot Cereal

5.3f ANALYSIS

Using SAS 9.4, we explored the different types of food in the groups listed above and whether or not they are causes of outbreaks. Both outbreak definitions were used in the hpgenselect procedure to conduct forward selection, backward elimination, and stepwise selection on Poisson models. The HPGENSELECT procedure performs model selection for generalized linear models (GLMs). It fits models for standard distributions in the exponential family, such as the normal, Poisson, and Tweedie distributions. In addition, PROC HPGENSELECT fits multinomial models for ordinal and nominal responses, and it fits zero-inflated Poisson and negative binomial models for count data. For all these models, the HPGENSELECT procedure provides forward, backward, and stepwise variable selection and includes Akaike's information criterion (AIC), a small-sample bias-corrected version of Akaike's information criterion (AICC), and the Schwarz Bayesian criterion (SBC) as selection criteria.[23]

5.4 RESULTS

The frequencies for the main food groups and all demographic variables created from the DHEC dataset are listed in Table 5.1. Our dataset contained 4058 *Salmonella* cases identified by DHEC from 2008-2015. Of those, DHEC identified 78 as being part of a statewide outbreak while using the FDA definition yielded 2565 cases as part of an outbreak. All variables listed in table 5.1 were considered for inclusion in the model.

Each outbreak definition was run through three Poisson model selection processes to see what foods are associated with outbreaks in South Carolina. The models went through forward selection, backward elimination, and stepwise selection.

5.4a DHEC DEFINITION

Using the DHEC definition of an outbreak, each selection model had different criteria in the final model (See Table 5.2). Forward selection chose the meat, dairy, greens, urban/rural, and Hispanic as the final criteria that cause outbreaks in South Carolina. Backward elimination chose meat, dairy, greens, and urban/rural while stepwise selection chose meat, and greens. Table 5.3 breaks down each parameter and the reasons it was kept in the model. In the forward selection model, meat consumption is 2.78 times as likely among outbreak cases, greens are 2.5 times as likely, and dairy is 0.52 times as likely when using the DHEC definition. None of the confidence limits cross 1 at α=0.05, so they are all significant associations.

All parameters are positively associated with an outbreak. At the very least, the DHEC definition believes that meat and greens are a reasonable indicator of being a part of a *Salmonella* outbreak in South Carolina. This definition also lets us know that there are key demographic characteristics that may lead to an outbreak, specifically urban/rural and Hispanic. Table 5.4 shows the forward model stratified by urban/rural and Hispanic and the backward model stratified by urban/rural. Both yielded the same results, meaning that Hispanic is only a marginal factor. When using the stratification, meat consumption is 2.9 times as likely among outbreak cases, greens are 2.51, and dairy is 0.55 times as likely when using the DHEC definition. Meat and greens are significant at $\alpha=0.05$, but the confidence limit for dairy crosses 1, so the result may not be significant.

5.4b FDA DEFINITION

Using the FDA definition, each selection model came back with the same criteria for the final model. Forward selection, backward elimination, and stepwise selection chose dairy as the final criteria that cause outbreaks in South Carolina. Tables 5.5 and 5.6 provide the information similar to the DHEC definition. Dairy is shown to be 1.3 times as likely to be associated with an outbreak in South Carolina. This result is significant at $\alpha=0.05$. Here we can say that whichever selection method is used, dairy is a food that may be associated with *Salmonella* outbreaks in South Carolina.

5.4c DIARY BREAKOUT

With dairy being the common group between both definitions, it was broken out into its components as listed in the methods. In the DHEC definition, sliced cheese and flavored milk powder were chosen to the model and in the FDA definition, only milk was chosen. Table 5.7 shows the relative risks for each by definition. With the DHEC definition, sliced cheese was 1.76 times as likely to cause an outbreak while using any type of flavored milk powder was 4.2 times as likely. Milk was 1.11 times as likely to cause an outbreak when using the FDA definition.

5.5 DISCUSSION

Foods associated with *Salmonella* outbreaks in South Carolina varied widely by definition used. Meat and greens were positively associated with outbreak cases while dairy was negatively associated with outbreak cases using the DHEC definition. Using the FDA definition, only dairy was positively associated with outbreak cases. The DHEC definition of an outbreak, with its 78 cases, yielded results that are more in line with what is known to cause foodborne outbreaks;[3] however, the FDA definition did identify dairy, and it is known that unpasteurized milk is a common cause of foodborne illness.[3] Breaking dairy out into its individual variables and running them through both definition models yielded very different results. Neither identified the same factors, which is interesting in itself. But, the similarity could come with the flavored milk powder in the DHEC model and milk in the FDA model. Many people mix flavored powder with milk itself, so the correlation could exist. There is a significant difference in the percentage of what is considered an outbreak when using the two definitions yielding different results when it came to modeling. The interesting comparison between the two definitions that dairy is the only food that is present in both models.

The results here fit with what are common causes of *Salmonella* outbreaks in the United States, but are only loosely related with what has caused outbreaks in South Carolina in the past 10 years.[3] Causes of outbreaks in SC have been poultry, cucumbers, cantaloupe, raw seafood, and unpasteurized milk.[25] Dairy in our study contains all forms of dairy, both pasteurized and unpasteurized, so without further breakdown into just raw milk, we can only infer a loose relation. The raw category we tested contains all forms of raw food, some of which are not associated with outbreaks in SC, thus furthering the need for further stratification.

5.5a STRENGTHS AND LIMITATIONS

In this study, the outbreaks were identified and people were surveyed without delay since the state mandates that probable *Salmonella* cases be identified within 3 business days. This helps to avoid possible information and selection bias. However, this study is still subject to bias. Since the laboratory test must identify the causative agent and report the illness to public health authorities, in this case DHEC Division of Acute Disease Epidemiology (DADE), the appropriate questionnaire may not be administered in a timely fashion. This could cause recall bias when the ill person is attempting to answer all questions. If they cannot recall what is eaten in the 7 days prior to becoming ill, they may answer "I don't know" or just guess to what they normally eat. This could over or underestimate the risk of certain foods eaten. We may not full know what exactly the patient may have eaten, leading to misclassification of foods eaten prior to illness and spurious associations could arise. However, there is no reason to believe that what is eaten during a normal week would be any different than what was eaten prior to illness and if something eaten was jarringly different than the norm, it is likely the person would remember. Thus, it is unlikely that these errors would have affected the risk.

Another potential issue with this study could be the way the outbreak definitions were calculated. The DHEC definition of an outbreak takes into account what is going on across the country as well as what is known about South Carolina in particular. For example, we may see a rise in cases in the summertime due to family picnics and people leaving food out on a table, causing it to warm in the sun and spoil. This most likely is not cause for an outbreak and thus DHEC would not identify it as such. Due to this, there is a small amount of cases (78) considered part of an outbreak. The very large number of potential cases identified by using the FDA definition suggests that it included a large number of false positive cases because foodborne outbreaks are not that common. The DHEC outbreak definition may therefore be more accurate even though it cannot be directly determined from these data.

The FDA definition was calculated using the primary pattern of the organism and the illness onset date. If 2 or more cases occurred in a calendar month, then they were identified as being part of an outbreak. The problem with this definition is that it does not look at cases ± 30 days, but rather calendar month. So, if a case occurs on May 31st and another on June 2nd, they would not be identified as part of the same outbreak. This could lead to an underestimate of cases in the outbreak, but after combing through the data, cases that should have been a “yes” to being included in the outbreak definition were and this did not lead to any misclassification.

5.6 CONCLUSION

The aim of this study was to evaluate the food causes of *Salmonella* outbreaks in South Carolinians using two standard definitions of an outbreak. Using two separate definitions yielded only dairy as a food common between both. In the DHEC definition, dairy was only 0.5 times as likely to cause an outbreak whereas in the FDA definition, that number jumped to 1.14 times as likely. This does tell us that unpasteurized milk and other dairy products can be a food that is associated with outbreaks in the state. Meat and dairy are also associated when using the DHEC definition.

The differences in the relative risks in the two definitions are also letting us know that living in an urban or rural environment and being Hispanic can also be associated with foodborne illness. However, due to the high number of differences in the definitions, further investigation will be needed to address the challenges of a rural southern state with high disease burden in locations with the highest number of cases.

5.7 WORKS CITED

1. Nyachuba DG. Food-borne illness: is it on the rise? Nutr Rev. 2010; 68: 257-269.
2. Centers for Disease Control and Prevention (CDC). Surveillance for foodborne disease outbreaks—United States, 2007. *Morbidity and Mortality Weekly Report.* 2010;59(31):973–979.
3. Centers for Disease Control and Prevention. Foodborne Illness, Foodborne Disease, (sometimes called "Food Poisoning"). http://www.cdc.gov/foodsafety/facts.html#what. Published 2014. Accessed 1 April 2014.
4. National Institute of Diabetes and Digestive and Kidney Diseases. Foodborne Illnesses. http://digestive.niddk.nih.gov/ddiseases/pubs/bacteria/. Published 2012. Accessed 1 April 2014.
5. US Department of Health and Human Services. Food Poisoning. http://www.foodsafety.gov/poisoning/. Published 2014. Accessed 1 April 2014.
6. Santos RL, Raffatellu M, Bevins CL, Adams LG, Tükel C, Tsolis RM, Bäumler AJ. Life in the inflamed intestine, *Salmonella* style. Trends Microbiol. 2009:17, 498–506.
7. Griffin AJ, McSorley SJ. Development of protective immunity to *Salmonella*, a mucosal pathogen with a systemic agenda. Mucosal Immunol. 2011;4:371–382.
8. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ. Typhoid fever. *N. Engl. J. Med.*. 2002; 347: 1770-1782.
9. Centers for Disease Control and Prevention (CDC). Multistate outbreaks of *Salmonella* infections associated with raw tomatoes eaten in restaurants—United States, 2005–2006. *Morbidity and Mortality Weekly Report.* 2007;56:909-911.
10. Centers for Disease Control and Prevention (CDC). Outbreak of *Salmonella* serotype Saintpaul infections associated with multiple raw produce items—United States, 2008. *Morbidity and Mortality Weekly Report.* 2008;57(34):929-934.

11. Centers for Disease Control and Prevention (CDC). Multistate outbreak of *Salmonella* infections associated with peanut butter and peanut butter-containing products—United States, 2008–2009. *Morbidity and Mortality Weekly Report.* 2009;58:85-90.
12. Centers for Disease Control and Prevention (CDC). Multistate outbreak of *Salmonella* serotype Tennessee infections associated with peanut butter—United States, 2006-2007. *Morbidity and Mortality Weekly Report.* 2007;56(21):521-524.
13. Olsen SJ, MacKinnon LC, Goulding JS, Bean NH, Slutsker L. Surveillance for food-borne-disease outbreaks—United States, 1993–1997. *Morbidity and Mortality Weekly Report Surveillance Summary.* 2000;49:1-62.
14. Rabsch W, Tschäpe H, Bäumler AJ. Non-typhoidal salmonellosis: emerging problems. *Microbes Infect*. 2001; 3(3):237-247.
15. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV. Food-related illness and death in the United States. *Emerg. Infect Dis.* 1999; 5: 607-625.
16. Chalker RB, Blaser MJ. A review of human salmonellosis: III. Magnitude of *Salmonella* infection in the United States. *Rev. Infect. Dis.* 1988;10:111-124.
17. Glynn MK, Bopp C, Dewitt W, Dabney P, Mokhtar M, Angulo FJ. Emergence of multidrug-resistant Salmonella enterica serotype typhimurium DT104 infections in the United States, *N. Engl. J. Med.* 1998;338:1333–1338.
18. Angulo FJ, Swerdlow DL. *Salmonella* enteritidis infections in the United States, *J. Am. Vet. Med. Assoc.* 1998;213:1729–1731.
19. Cohen ML, Tauxe RV. Drug-resistant *salmonella* in the United States: an epidemiologic perspective, *Science.* 1986;234:964–969.
20. St. Louis ME, Morse DL, Potter ME, DeMelfi TM, Guzewich JJ, Tauxe RV, Blake PA. The emergence of grade A eggs as a major source of *Salmonella* enteritidis infections. New implications for the control of salmonellosis. *JAMA* 1988;259:2103–2107.
21. Rodrigue DC, Tauxe RV, Rowe B. International increase in *Salmonella* enteritidis: a new pandemic? *Epidemiol. Infect.* 1990;105:21–27.
22. U.S. Food and Drug Administration. What You Should Know About Government Response to Foodborne Illness Outbreaks. http://www.fda.gov/Food/ResourcesForYou/Consumers/ucm180323.htm. Published 2014. Accessed 20 October 2014.
23. SAS Institute. High-Performance Variable Selection for Generalized Linear Models: PROC HPGENSELECT. https://support.sas.com/rnd/app/examples/stat/hpgenselect/hpgenselect.pdf Published 2015. Accessed 20 February 2016.
24. U.S. Department of Agriculture Economic Research Service. South Carolina - Rural Definitions: State Level Maps. http://www.ers.usda.gov/datafiles/Rural_Definitions/StateLevel_Maps/SC.pdf. Published 2015. Accessed 15 Nov 2015.
25. Marler Clark. Foodborne Illness Outbreak Database. http://outbreakdatabase.com/site/search/?outbreak=&vehicle=&organism=Salmonella&month=&year=&state=58&country=&x=49&y=13. Accessed 10 Apr 2016.

5.8 TABLES

TABLE 5.1 STUDY POPULATION CHARACTERISTICS OF THE 4058 *SALMONELLA* CASES

Value	Total		Outbreak*		Outbreak**		Gender		Hispanic		Race		Urban/Rural	
	N (YES)	%	N (YES)	%	N (YES)	%	N (Male)	%	N (YES)	%	N (White)	%	N (Urban)	%
Poultry	1747	43.05%	44	1.08%	1166	28.73%	856	21.09%	76	1.87%	1302	32.08%	1518	37.41%
Meat	1595	39.31%	52	1.28%	1069	26.34%	792	19.52%	59	1.45%	1234	30.41%	1377	33.93%
Pork	1499	36.94%	48	1.18%	1003	24.72%	769	18.95%	61	1.50%	1107	27.28%	1299	32.01%
Seafood	691	17.03%	25	0.62%	459	11.31%	317	7.81%	23	0.57%	481	11.85%	600	14.79%
Eggs	1316	32.43%	33	0.81%	863	21.27%	652	16.07%	60	1.48%	1006	24.79%	1127	27.77%
Dairy	2006	49.43%	50	1.23%	1351	33.29%	965	23.78%	86	2.12%	1529	37.68%	1745	43.00%
Fruit	1715	42.26%	44	1.08%	1140	28.09%	808	19.91%	82	2.02%	1285	31.67%	1493	36.79%
Vegetables	1491	36.74%	46	1.13%	995	24.52%	678	16.71%	63	1.55%	1157	28.51%	1283	31.62%
Greens	1031	25.41%	41	1.01%	701	17.27%	462	11.38%	44	1.08%	821	20.23%	891	21.96%
Snacks	1765	43.49%	52	1.28%	1179	29.05%	858	21.14%	76	1.87%	1345	33.14%	1530	37.70%
Nuts	1224	30.16%	33	0.81%	815	20.08%	595	14.66%	38	0.94%	967	23.83%	1075	26.49%
Grains	1466	36.13%	41	1.01%	961	23.68%	721	17.77%	57	1.40%	1117	27.53%	1279	31.52%
Raw	710	17.50%	25	0.62%	479	11.80%	337	8.30%	29	0.71%	558	13.75%	596	14.69%
Frozen	1325	32.65%	32	0.79%	865	21.32%	642	15.82%	51	1.26%	1028	25.33%	1152	28.39%
Pre-Pkg	2020	49.78%	55	1.36%	1340	33.02%	970	23.90%	84	2.07%	1542	38.00%	1741	42.90%

* Using DHEC definition
** Using FDA definition

TABLE 5.2 MODEL SELECTION – DHEC DEFINITION

Selection Type	Variables Selected
Forward	meat, dairy, greens, urban/rural, hispanic
Backward	meat, dairy, greens, urban/rural
Stepwise	meat, greens

TABLE 5.3 RELATIVE RISKS FOR EACH VARIABLE SELECTED TO THE MODEL – DHEC DEFINITION

	Forward				Backward				Stepwise			
Parameter	***Relative Risk***	***Standard Error***	***Confidence Limits***		***Relativ e Risk***	***Standar d Error***	***Confidence Limits***		***Relative Risk***	***Standard Error***	***Confidence Limits***	
Meat	2.782	0.945	1.4298	5.412	2.840	0.974	1.4503	5.562	2.078	0.5813	1.201	3.5957
Green	2.499	0.702	1.440	4.3342	2.497	0.705	1.4356	4.344	2.214	0.5848	1.3197	3.7157
Dairy	0.515	0.167	0.272	0.9738	0.507	0.166	0.2664	0.9639				
Urban/Rural	2.652	1.135	1.146	6.1347	2.600	1.113	1.1238	6.0144				
Hispanic	0.948	0.051	0.854	1.0523								

TABLE 5.4 RELATIVE RISKS FOR EACH VARIABLE SELECTED TO THE MODEL STRATIFIED BY URBAN/RURAL AND HISPANIC – DHEC DEFINITION

Parameter	***Relative Risk***	***Standard Error***	***Confidence Limits***	
Meat	2.898	0.9998	1.4737	5.699
Green	2.507	0.709	1.440	4.3641
Dairy	0.547	0.179	0.287	1.0401

TABLE 5.5 MODEL SELECTION – FDA DEFINITION

Selection Type	Variables Selected
Forward	dairy
Backward	dairy
Stepwise	dairy

TABLE 5.6 RELATIVE RISKS FOR EACH VARIABLE SELECTED TO THE MODEL – FDA DEFINITION

Parameter	*Relative Risk*	*Standard Error*	*Confidence Limits*	
Dairy	1.138	0.045	1.0535	1.230

TABLE 5.7 RELATIVE RISKS FOR DAIRY VARIABLES CHOSEN TO EACH DEFINITION'S MODEL

	DHEC				**FDA**			
Dairy Group Components	***Relative Risk***	***Standard Error***	***Confidence Limits***		***Relative Risk***	***Standard Error***	***Confidence Limits***	
Sliced Cheese	1.759	0.4361	1.082	2.859				
Flavored Milk Powder	4.195	2.176	1.518	11.597				
Milk					1.111	0.045	1.0261	1.2028

CHAPTER 6

DIETARY INTAKE AND FOOD CAUSES OF FOODBORNE ILLNESS OF SOUTH CAROLINIANS – EVALUATION OF THE SOUTH CAROLINA DEPARTMENT OF HEALTH AND ENVIRONMENTAL CONTROL FOOD EXPOSURES SURVEY[3]

[3] Alianell, A.T., Merchant, A., McLain, A., Brenner, E., and D. Giurgiutiu. To be submitted to *American Journal of Epidemiology*.

6.1 ABSTRACT

Dietary patterns are associated with mortality from all causes, which raises the need for public health approaches to ensure that healthy food options are available, accessible, and affordable for all South Carolinians. We aimed to characterize the nutrition and dietary intake of South Carolina residents using the South Carolina Department of Health and Environmental Control Food Exposures Survey and also see what foods eaten may be associated with foodborne outbreaks. Using the data provided for 875 individuals across South Carolina, over 90% of participants answered yes to eating poultry, meat, dairy, fruit, vegetables, raw food, pre-packaged foods, and frozen items. Those who identified as Hispanic had the most significant associations with the food categories when looking at the state as a whole as well as controlling for urban and rural environments. Eggs were the least eaten food when looked at by demographic characteristics, yielding a significant association with age and income. Due to the high number of differences in the population, further investigation will be needed to address the challenges of accessibility and affordability to different food options in South Carolina.

6.2 INTRODUCTION

Dietary patterns are associated with mortality from all causes, coronary heart disease, cardiovascular diseases, and cancer.[1] A healthy eating pattern has been associated with a reduced mortality risk and reduced obesity.[2,3] Diets that consist of a high intake of vegetables, legumes, fruits, nuts, cereals, and a high intake of olive oil but a low intake of saturated lipids, a moderately high intake of fish, a low-to-moderate intake of dairy products, a low intake of meat and poultry, and a regular but moderate intake of ethanol (wine) are what help reduce the risk of obesity and overall mortality risk.[4] South Carolina now has the 10th highest adult obesity rate in the nation reports the *The State of Obesity: Better Policies for a Healthier America*.[5] According to the Centers for Disease Control and Prevention (CDC), 66.9% of adults in South Carolina are overweight (BMI ≥ 25) and 31.5% are obese (BMI ≥ 30).[6,7] Only 23.3% and 22.9% of adults have reported having consumed 2 or more servings of fruits and 3 or more servings of vegetables at the recommended levels, respectively in the state.[6,8] Adolescents have unhealthy dietary behavior as well, with 74.8% eating fruits or drinking 100% fruit juice less than 2 times per day and 91.2% eating vegetables less than 3 times per day. 33.2% consume sugar-sweetened beverage consumption at least once per day.[6] A need for public health approaches is sought after to ensure that healthy food options are available, accessible, and affordable for all South Carolinians.

Studies have suggested that race/ethnicity and socioeconomic status (SES) influence food choices and dietary compliance, which may further modify the associations between diet and health outcomes.[9-11] African-Americans and low-SES adults have been shown to have more limited access to supermarkets and healthy foods, tend to consume more energy-dense and nutrient-poor foods, and to have generally lower diet quality scores, when compared with white adults.[12-15] Improvements in the diets of South Carolinians can potentially improve the risks for obesity and disease.

To assess the food preferences South Carolinians consume, the South Carolina Department of Health and Environmental Control (DHEC) commissioned the University of South Carolina's Institute for Public Service and Policy Research (IPSPR) to conduct a telephone survey with a representative sample of the state's population two years of age or older.[16] The questionnaire used in this survey was initially designed by DHEC staff and based on the hypothesis-generating questionnaire that is used by DHEC when it encounters cases of *Salmonella* and Shiga toxin-producing *E. coli (*STEC). The questionnaire was modified into the Food Exposures Survey to address the questions of foods eaten without being in relation to an outbreak or confirmed case of foodborne illness. This helps to aid in the assessment of overall eating patterns and not just those around the time the study participant may have gotten ill. Each of the questions was administered to a representative sample of the target population, aimed at representing the entire state. Since this study was designed based on the hypothesis generating questionnaire, it is possible to take the dietary patterns and the demographic characteristics of South Carolinians gathered and determine whether or not they may lead to foodborne outbreaks.

Foodborne illness constitutes a serious public health threat with the Centers for Disease Control and Prevention (CDC) estimating that 76 million foodborne illnesses, including 325,000 hospitalizations and 5,000 deaths, occur in the United States each year and anyone is susceptible to becoming sick.[17,18] Many pathogens, like bacteria, viruses, and parasites, can contaminate foods or beverages, leading to many different foodborne infections.[19] One of the most common causes of foodborne illness is *Salmonella,* a Gram-negative facultative intracellular pathogen that causes a spectrum of clinical diseases depending on the serotype of the infecting bacteria and the susceptibility of the host.[20,21] The most common symptoms of *Salmonella* infection are nausea, vomiting, abdominal cramps, and diarrhea, but most healthy individuals will recover without the need for treatment in about a week.[19,22,23]

The aim of the current study is to first, characterize the nutrition and dietary intake of South Carolina residents using the South Carolina Department of Health and Environmental Control Food Exposures Survey and second, to create a baseline of foods eaten normally in South Carolina to aid in foodborne outbreak investigations in South Carolina. We will look at the demographic, social, and geographic determinants of dietary groups in the South Carolina population as well as looking for any associations between the dietary patterns and frequency of eating pre-packaged foods or fresh foods. The dietary groups are categorized to be consistent with the healthy eating patterns listed above. This information will help characterize one of the important factors contributing to health in a southern state with a high obesity rate as well as show what foods associated with outbreaks are being eaten by certain demographic populations.

6.3 METHODS

6.3a STUDY PARTICIPANTS

875 participants were included in this analysis. Data collection was done via telephone from July 27-August 30, 2012 from 9:00 AM to 9:30 PM Monday through Friday, from 10:00 AM to 4:00 PM on Saturday, and 3:00 PM to 8:00 PM on Sunday.[16] A dual sampling frame approach was used in selecting study participants. One sampling set was based on landline telephone exchanges and the second on cell phone telephone numbers. For the landline component, respondents were selected from a random sample of households with telephones in the state. Respondents in the cell phone sample were randomly selected from a list of cell phone exchanges in South Carolina. All phones with a South Carolina exchange were eligible for the study and all counties are included in the

sample. The survey interviewers called each of these numbers. Numbers that were found to be businesses, institutions, not-in-service, or otherwise not assigned were ineligible for the survey.

The remaining numbers that were not excluded were called, which resulted in contact in both the landline component and the cell phone component of the study. When contact was made with a residence in the landline component, a participant two years of age or older was randomly chosen from the occupants of the household. If the selected participant was between the ages of 2 and 11, an adult in the household was asked to be a proxy to answer the questions for the child. Participants aged 12 or older were interviewed directly about their food exposure experiences. Proxy interviews were conducted for participants between the ages of 12 and 17 if an adult in the household did not want the selected child to participate.[16]

6.3b CONFIDENTIALITY

Information collected and used in this study does not contain any personal information and the investigators using the data will not have access to any of the personal information that may be on file at DHEC. This study falls under non-human subjects research according to the application that was submitted, reviewed, and approved by both the University of South Carolina Institutional Review Board (IRB) and the DHEC IRB. The data used in this study is data previously collected by DHEC's Division of Acute Disease Epidemiology (DADE). A unique key has been assigned to each person before the start of this analysis. No individual will be identified in any publications resulting from this study.

When contact was made with an individual in the cell phone component, they were asked a series of questions to determine eligibility, including confirming that the number reached was for a cell phone, that the individual who answered was 12 years of age or older, and that they were a resident of South Carolina. If all criteria were met, they continued with the survey. If an individual that was reached on a cell phone also had a landline telephone and received less than 90% of their calls on their cell phone, they were considered ineligible and thus not interviewed.

6.3c ASSESSMENT OF OUTCOME (FOOD GROUP DEFINITIONS)

The food preferences questionnaire used in this survey was based on the hypothesis-generating questionnaire that is used by DHEC when it encounters a confirmed case of *Salmonella* or STEC. The goal of this survey was to limit the amount of time required to complete the interview to around twenty minutes. Given this goal, all items contained in the hypothesis generating questionnaire could not be included in this survey and thus two versions of the questionnaire were developed. Some questions appear on both forms, while others appear on only one, with 81% of the questions being the same on both versions. On version 2, some questions were broken out that were combined in version 1, but those were then combined in analysis since it was just the manner in which it was asked that differed. (A copy of the questionnaires is included in Appendix C.) The demographic questions and other technical aspects of the questionnaire were designed by University of South Carolina's Institute for Public Service and Policy Research (IPSPR) staff in conjunction with DHEC.[16] Each of these questionnaires was administered to a representative sample of the target population.

Before the questionnaires were finalized they were pretested within IPSPR to determine whether or not the questions could be easily understood by respondents, if the order of the questions seemed logical to the interviewers and respondents, or if they contained other identifiable weaknesses. Problems were detected and corrected. No major problems persisted into the actual conduct of the survey. With the survey, version 1 contained 145 questions and version 2 contained 147. Each food question was asked for a yes/no/don't know response based on what was eaten in the past 7 days. Each section was asked with specifics in mind, for example, with vegetables, the following was read prior to the respondent answering:

> "The next questions are about fresh vegetables you (your child) might have eaten raw or uncooked in the past seven days. These foods could have been eaten either in the home or away from home. This does not include canned items, but these foods could have been eaten alone or as part of a dish. We are only interested in vegetables that are not grown at home. In the past seven days, did you (your child) eat any:"[16]

With the demographics, the answers were recorded differently than the food, with the exception of Hispanic origin, which was still yes/no/don't know. Age and number of residents were coded as an exact integer, with number of residents being split into children and adults. Exact zip code was entered to extract the county of residence while income and race were broken down into specific categories for ease of analysis. Gender was recorded as male or female only.

6.3d CATEGORIZATION OF FOODS

Foods eaten by participants were grouped into specific categories based on the way the questionnaire was broken up for ease of analysis. The groups are poultry, meat, pork, seafood, egg, dairy, fruit, vegetables, greens, snacks, nuts, grains, raw foods, frozen foods, and prepackaged foods that are not frozen. The raw, frozen, and prepackaged categories may contain foods that are already in the other categories. They were split for further analysis.

The food grouping categories are listed below.

- Poultry: Whole or Cut Chicken Pieces/Parts; Ground Chicken; Breaded Chicken Products, such as chicken tenders and the like; Stuffed, Frozen Chicken Products, such as Chicken Kiev and the like; Other Frozen Chicken, Duck, Game Hen, or Squab; Whole or Cut Turkey Pieces or Parts; Ground Turkey
- Meat: Beef Steaks or Roasts; Pre-made or Pre-formed Hamburger Patties at Home; Fresh Hamburger Patties at Home; Any Other Ground Beef; Lamb; Store-bought, Dried Meat Strips or Jerky; Any Other Italian Meats; Bologna, Pastrami, or Corned Beef; Pre-packaged Deli Meats; Any Other Deli-sliced Meats not Pre-packaged
- Pork: Ground Pork; Pulled Pork Barbecue; Other Pork; Bacon; Sausage; Hot Dogs, Corn Dogs, Polish Sausage, Kielbasa, or similar foods; Pepperoni

- Seafood: Fresh or Fresh-Frozen fish; Smoked or Dried Fish; Shrimp or Prawns; Crab, Lobster, or Crayfish; Oysters; Clams, Mussels, Scallops, or Other Shellfish; Sushi with Raw Fish or Seafood; Frozen Fish Products, such as Fish Sticks, Fish Nuggets, and the like
- Egg: Eggs or Dishes Containing Eggs; Anything Made with Raw Eggs
- Dairy: Milk; Raw or Unpasteurized Milk; Any Other Dairy (Soy or Almond Milk); Ice Cream; Frozen Yogurt; Yogurt Drinks; Other Yogurt; Pre-packaged Shredded Cheese; Processed Sliced Cheese; Block Cheese such as Cheddar, Swiss, Colby, and the like; String Cheese; Cottage Cheese; Feta; Blue Veined Cheese (Gorgonzola or Bleu); Fresh or Dried Parmesan or Similar Cheese; Cheese from Raw or Unpasteurized Milk (Homemade or Farm-fresh); Queso Fresco or Queso Blanco; Homemade Mexican-Style Soft Cheese; Dry Buttermilk; Flavored Milk Powder; Other Powdered Milk
- Fruit: Apple; Lemon/Lime; Strawberry; Raspberry; Blueberry; Cherry; Any Other Fresh Berry; Cantaloupe; Honeydew; Watermelon; Precut Melon or Melon Salad; Any Other Melon; Pineapple; Mango; Any Other Tropical Fruit; Any Other Dried Fruit; Apple Juice not from Concentrate; Orange Juice not from Concentrate; Grape; Pear; Peach; Nectarine; Apricot; Plum; Orange; Grapefruit; Tangerine; Any Whole or Shredded Coconut; Raisins
- Vegetables: Red Round Tomato at Home; Roma Tomato at Home; Cherry Tomato at Home; Grape Tomato at Home; Tomatoes on the Vine at Home; Any Other Fresh Tomato at Home; Any Tomato Away from Home; Fresh Salsa or Pico de Gallo (not from a jar); Guacamole; Alfalfa Sprouts; Other

Sprouts; Fresh Chili Peppers (Serrano, Poblano, or the like); Mini Carrots; Any Raw Onion; Raw Green Onion or Scallion; Cucumber, Zucchini, or Squash; Bell Pepper; Celery; Other Fresh Carrots; Any Other Root Vegetables; Fresh, Raw Peas; Broccoli; Cauliflower; Fresh or Dried Mushrooms; Any Frozen Vegetables

- Greens: Pre-packaged Greens; Loose Greens; Pre-packaged Iceberg; Loose Iceberg; Pre-packaged Romaine; Loose Romaine; Pre-packaged Spinach; Loose Spinach; Cabbage; Other Greens; Basil; Cilantro; Other Herbs
- Snacks: Frozen Pot Pie; Pre-packaged Peanut Butter Crackers; Chips or Pretzels; Pre-packaged Crackers, Cookies, or Snack Cakes; Chocolate or Candy Containing Chocolate; Frozen Pizza; Frozen Mexican Food; Frozen Snacks; Frozen Breakfast; Frozen Vegetarian; Frozen Bagged Meal; Frozen Dinner; Fruit Roll-up; Trail Mix; Powdered Nutrition Supplement
- Nuts: Any Peanut Butter; Creamy Peanut Butter; Crunchy Peanut Butter; Foods Containing Peanut Butter; Nut Butter; Peanuts; Almonds; Walnuts; Cashews; Pistachios; Hazelnut; Sunflower Seeds; Sesame Seeds; Tahini; Hummus
- Grains: Granola Bars, Breakfast, Power, or Protein Bars; Cereal; Hot Cereal

6.3e OTHER COVARIATES

Other variables used in analysis included demographic groups broken down as urban or rural environment, age categories, gender, Hispanic ethnicity, income, and race. Urban or rural was categorized by using the county of residence from the questionnaire and then the rural definition based on the Office of Management and Budget (OMB) metropolitan statistical areas and their categorization of each county in the state of South Carolina.[24]

Gender was classified as 0 for female and 1 for male for ease of analysis. Age was categorized as an integer with the option of being missing if the participant refused. Race was put into categories of Caucasian, African American, other, and unknown/refused. Hispanic was asked as whether or not the person being interviewed identified as Hispanic with the answers being categorized as 0 for no, 1 for yes, 88 for refused, 99 and missing as unknown. Income was classified into categories of <$25,000, $25-49,999, $50-99,999, ≥ $100,000, and blank for refused or missing.

For calculations of relative risk, each demographic was made into binary variables. Gender, Hispanic, and urban remained as they were with female/male, no/yes, and urban/rural, respectively. Age changed to < 45 and ≥ 45 years while race changed to Caucasian and other. Income became under $50,000 and $50,000 or greater.

If questions in the questionnaire specifically asked for frozen foods, raw foods, or pre-packaged foods, they were placed into groups as well. These groups overlap with the food groups themselves and were created as such:

- Raw: Sushi with Raw Fish or Seafood; Anything Made with Raw Eggs; Raw or Unpasteurized Milk; Cheese from Raw or Unpasteurized Milk (Homemade or Farm-fresh); Red Round Tomato at Home; Roma Tomato at Home; Cherry Tomato at Home; Grape Tomato at Home; Tomatoes on the Vine at Home; Any Other Fresh Tomato at Home; Any Tomato Away from Home; Fresh Salsa or Pico de Gallo (not from a jar); Guacamole; Pre-packaged Greens; Loose Greens; Pre-packaged Iceberg; Loose Iceberg; Pre-packaged Romaine; Loose Romaine; Pre-packaged Spinach; Loose Spinach; Other Greens; Any Raw Onion; Raw Green Onion or Scallion; Fresh, Raw Peas; Cabbage

- Frozen: Stuffed, Frozen Chicken Products, such as Chicken Kiev and the like; Other Frozen Chicken, Duck, Game Hen, or Squab; Fresh or Fresh-Frozen fish; Frozen Fish Products, such as Fish Sticks, Fish Nuggets, and the like; Ice Cream; Frozen Yogurt; Frozen Pot Pie; Frozen Vegetables; Frozen Pizza; Frozen Mexican Food; Frozen Snacks; Frozen Breakfast; Frozen Vegetarian Meal; Frozen Bagged Meal; Frozen Dinner
- Pre-Packaged: Pre-Packaged Deli Meat; Pre-made or Pre-formed Hamburger Patties at Home; Pre-packaged Shredded Cheese; Processed Sliced Cheese; Pre-packaged Greens; Pre-packaged Iceberg; Pre-packaged Romaine; Pre-packaged Spinach; Pre-packaged Peanut Butter Crackers; Pre-packaged Crackers, Cookies, or Snack Cakes; Granola Bars, Breakfast, Power, or Protein Bars; Cereal; Hot Cereal; Chips or Pretzels; Chocolate or Candy Containing Chocolate; Fruit Roll-up; Trail Mix

6.3f ANALYSIS

Since there were two versions of the questionnaire, the data for those items that were included in both forms of the questionnaire were combined. For example, question 1 on both versions asked about the consumption of whole or cut chicken pieces/parts in the past 7 days. For the items on “any other ground beef” (Version 1, Q12; Version 2, Q9) “any other pork product” (Version 1, Q18; Version 2, Q11), “pre-packaged deli meats”

(Version 1, Q24; Version 2, Q12), and "other fresh tomatoes at home" (Version 1, Q77; Version 2, Q62), the context in which the questions were asked was slightly different, but the results were grouped for ease of analysis since the answers were still yes/no/don't know.[16] To avoid biasing the sample in favor of households that can be reached on multiple landline telephone numbers, each case from the landline sample was weighted inversely to its probability of being included in the sample and adjusted for differences in probability of selection due to the number of individuals living in the household. The data were also weighted to correct any potential biases in the sample on the basis of age, race, and sex. The data from the landline component were first weighted to adjust for households that can be reached on more than one telephone number to correct overrepresentation of that household. They were then weighted due to the fact that it was a household that could have multiple members living there and not an individual respondent like the cell phone component. The final part of weighting dealt with the underrepresentation of certain demographic variables as assessed by population estimates from the US Census Bureau. The final variable used in analysis was CENSWT2.

Using SAS 9.4, we explored the different types of food in the groups listed above and whether or not certain demographic characteristics had an effect on the foods eaten as well as whether or not certain foods cause outbreaks in South Carolina. Both versions of the questionnaire were grouped together to use the population as a whole. The food groups were used in the hpgenselect procedure to conduct forward selection, backward elimination, and stepwise selection on Poisson models. If all models turn out to produce the same results, stepwise selection will be used. After the model selection was conducted, relative risk was calculated on each demographic variable selected with an $\alpha=0.05$.

6.4 RESULTS

There were 445 participants (324 landline, 121 cell) that answered version 1 of the questionnaire and 430 (311 landline, 119 cell) that answered version 2 for a total of 875 participants from the main survey period of July 27 to August 30, 2012. The average age of the participants was 41 years old with 22% having said their income was between $50,000 and $99,999. Other demographics calculated showed that 51% were female, 3.5% were Hispanic, 62% identified as Caucasian, and 74% said they lived in an urban setting. The population characteristics of the participants are listed in Table 6.1. The total yes responses for poultry, meat, dairy, fruit, vegetables, raw food, pre-packaged goods, and frozen food are of note in table 6.1 having over a 90% yes response rate. All demographic variables listed were used in calculating relative risk.

Table 6.2 shows the relative risk of each type of food by each demographic variable. Females were 0.94 times as likely to consume meat, 0.86 times as likely to eat pork, and 1.05 times as likely to have vegetables relative to males. When it came to age, those under 45 were 1.07 times as likely to have meat, 1.1 times as likely to eat pork, 1.06 times as likely to consume eggs, 1.2 times as likely to eat snacks, 1.1 times as likely to have nuts, 1.3 times as likely to eat grains, and 1.03 times as likely to eat pre-packaged products as compared to those 45 and older. Those who identified as Hispanic were 1.08 times as likely to eat poultry, 1.1 times as likely to eat meat, fruit, nuts, raw food and vegetables, and 1.2 times as likely to consume greens relative to those who do not identify as Hispanic.

The only categories that were not significant when looking at Caucasians relative to other races were pork, eggs, and fruit. They were 1.04 times as likely to eat poultry and 1.05 times as likely to eat meat while only 0.75 times as likely to consume seafood. Caucasians were also 1.05 times as likely to have any dairy products, 1.1 times as likely to eat vegetables and snacks. When it came to eating greens, nuts, grains, raw food, frozen items, and pre-packaged products, they were 1.09, 1.23, 1.22, 1.05, 1.05, and 1.03 times as likely relative to other races, respectively. Those making under $50,000 were 1.04 times as likely to consume meat, 1.05 times as likely to eat eggs, 1.07 times as likely to eat snacks, 1.06 times as likely to eat nuts, and 1.03 times as likely to eat any raw food relative to those making $50,000 and over. The results when looking at an urban setting yielded significance in poultry (1.08), meat (1.03), dairy (1.04), raw food (1.06), and frozen food (1.06) relative to those who live in a rural environment.

Given that South Carolina has a very diverse urban and rural population, with 21 counties considered urban and 25 considered rural, each relative risk in Table 6.2 was recalculated to control for urban vs. rural.[27] The results are presented in Table 6.3 for urban and 6.4 for rural. When controlling for an urban environment, females were 0.86 times as likely to eat meat and 0.80 times as likely to eat meat relative to males. Those under 45 were 1.07 times as likely to eat meat, 1.15 times as likely to consume snacks and 1.26 times as likely to eat grains, but only 0.79 times as likely to consume seafood and 1.08 times as likely to eat eggs relative to those 45 and older when controlling for an urban setting. The results when looking at those of Hispanic origin and controlling for urbanity yielded significance in poultry (1.05), meat (1.09), dairy (1.02), fruit (1.04), vegetables (1.09), snacks (1.12), nuts (1.11), raw food (1.05), frozen items (1.07), and pre-packaged products

(1.02) relative to those who do not identify as Hispanic. Caucasians were only 0.75 times as likely to consume seafood when compared to other races in an urban environment; but, Caucasians were 1.03 times as likely to consume dairy, 1.08 times as likely to eat vegetables, 1.09 times as likely to eat snacks, 1.22 times as likely to eat nuts, and 1.23 times as likely to consume grains relative to other races in an urban environment. People who make under $50,000 were 1.08 times as likely to eat nuts, 1.23 times as likely to have grains, and 1.04 times as likely to eat raw food relative to those who make $50,000 or more in an urban environment.

When controlling for a rural setting, females were 1.14 times as likely to eat poultry, 1.11 times as likely to eat vegetables, and 1.28 times as likely to have grains but only 0.91 times as likely to eat meat relative to males. Those under 45 years of age in a rural environment were 1.19 times as likely to eat pork and snacks, 1.16 times as likely to have nuts, 1.46 times as likely to have grains, and 1.07 times as likely to eat pre-packaged items relative to those 45 and older. People who identified as Hispanic and live in a rural environment were 1.15 times as likely to eat poultry, 1.12 times as likely to have meat, 1.93 times as likely to eat seafood, 1.08 times as likely to consume fruit, 1.14 times as likely to eat vegetables, 1.29 times as likely to eat greens, 1.15 times as likely to have any nuts, 1.58 times as likely to consume grains, and 1.11 times as likely to have raw food when compared to people who do not identify as Hispanic. The results for Caucasians in a rural setting yielded significance in poultry, dairy, vegetables, snacks, nuts, and pre-packaged foods relative to other races. They were 1.19 times as likely to consume poultry, 1.09 times as likely to have any dairy, 1.18 times as likely to have vegetables, 1.26 times as likely to consume any nuts, and 1.05 times as likely to have any pre-packaged foods.

People who make under $50,000 were 1.18 times as likely to eat pork and 1.15 times as likely to eat snacks but only 0.90 times as likely to consume fruit relative to those who make $50,000 or more when controlling for rurality. Hispanics were 1.14 times as likely to eat eggs in a rural environment.

Each food group was run through three Poisson model selection processes to find demographics associated with eating certain types of food. The models went through forward selection, backward elimination, and stepwise selection. Each model came back with the same criteria for the final model, so the stepwise model was chosen. Only seafood, snacks, nuts, and grains had associations with demographic characteristics (Table 6.5). Seafood and nuts were associated with race, snacks with age, and grains with age and race. Table 6.6 shows the results of the variables selected to the seafood model. The study participants that identified as Caucasian were 0.75 times as likely to eat seafood when compared to other races. Table 6.7 shows the results of the snack model. Those under 45 were 1.16 times as likely to consume snacks compared to those 45 and older. Table 6.8 shows that Caucasians are 1.23 times as likely to consume nuts compared to other races. The results for the variables selected to the grains model are shown in table 6.9. Those under 45 are 1.33 times as likely to consume grains compared to those 45 and older and Caucasians are 1.23 times as likely to consume grains when compared to other races. All results listed were significant at $\alpha=0.05$.

6.5 DISCUSSION

Foods eaten by South Carolinians varied widely when looking at the demographic characteristics presented in this study. Over 90% of participants in the study answered yes to eating poultry, meat, dairy, fruit, vegetables, raw food, pre-packaged foods, and frozen items when using the census weight. Those who identified as Hispanic had the most significant associations with the food categories when looking at the state as a whole as well as controlling for urban and rural environments. Eggs were the least eaten food when looked at by demographic characteristics, yielding a significant association with age and income. It is interesting to note too that younger people are more likely to consume eggs than those that are older in both the state as a whole and in an urban setting, but that there is no significant association between eggs and age in a rural setting.

It is should be noted that only seafood, snacks, nuts, and grains had an association with demographic characteristics. Seafood had a negative association with race while snacks, nuts, and grains had positive associations with age, race, and age/race, respectively. The results could aid in foodborne outbreak investigations by providing a baseline of commonly eaten foods amongst certain demographic groups. It could help in identifying an actual outbreak since we know that certain food groups are more associated with certain demographics. For example, if there is suspicion of a nut outbreak, we should note that Caucasians are more likely to consume nuts on a regular basis, and thus their hypothesis-generating questionnaires should be looked at a little more closely. The results could also aid in the follow-up investigations by knowing who to target first in the suspected outbreak. We have also seen that foods commonly eaten like meat and poultry have no differences among demographic characteristics, which is to be expected.

6.5a STRENGTHS AND LIMITATIONS

The interviews for this study were conducted over a one month period in the summer with a questionnaire that is based on the hypothesis generating questionnaire that is used by DHEC when it encounters a confirmed case of *Salmonella*, STEC, or *Listeria*. Given that this questionnaire was developed for another purpose and is rather lengthy, this could have posed a problem for this study. However, the questionnaire was changed given the goal of limiting the time required to complete to 20 minutes and some questions were dropped. It was also made into an easier answer format of yes/no/don't know, so this should not have been a problem.

When the survey was conducted, both landlines and cell phone exchanges in South Carolina were used. There is no issue with the landline as you must have a South Carolina area code to have a phone in your house, but issues can exist with the cell phone exchanges chosen. A representative sample of the South Carolina population may not be achieved since only SC area codes were also chosen for cell phones. If someone moved to South Carolina, such as a student moving here for college, they may not have chosen to change their cell phone number and thus have an exchange from another state. This removes a part of the population from being eligible for this study before any participants are chosen. To remediate this problem, the census weight was used in all calculations to create a more representative target population.

This survey, like all surveys, has the potential for sampling error given that only a part of the population of the state was interviewed and not all residents of South Carolina participated. For the questions that were answered by at least 800 respondents, the potential error is very low, but those answered by significantly less than 800 respondents have the potential for a larger variation than those for the entire sample. To help reduce this potential error, foods and demographics were grouped to reduce the variation.

6.6 CONCLUSION

The aim of the current study was to characterize the nutrition and dietary intake of South Carolina residents using the DHEC Food Exposures Survey. The foods eaten by South Carolinians varied widely when participants were placed into demographic groups. By using a representative sample that contains geographic and socio-demographic diversity, we can use the information to potentially affect change in the food and dietary programs available to South Carolinians.

We aimed to look at the demographic, social, and geographic determinants of dietary patterns in the South Carolina population that may or may not have been investigated for foodborne illness outbreaks as well as looking for any associations between the dietary patterns and frequency of eating pre-packaged foods or fresh foods. This information will address the challenges of a rural southern state with a high obesity rate by using a representative sample that contains geographic and socio-demographic diversity and using said information to help affect change in the programs available. Participants that identified as Hispanic and Caucasians consumed the highest amount of fresh food while those under 45 years of age and Caucasians consumed the most pre-packaged food. Caucasians and those that live in an urban environment eat the most frozen food relative to other races and those in a rural environment, respectively. When controlling for an urban or rural setting, the results changed to show that while the fresh food demographics did not change, the frozen and pre-packaged relative risks did. It was those who identified as Hispanic that were more likely to consume frozen or pre-packaged items relative to those that did not identify as Hispanic. The rural setting more closely resembled the state as a whole.

Seafood, snacks, nuts, and grains had associations with demographic characteristics, which could aid in foodborne investigations by providing more knowledge about what is eaten in the state as a whole. We can target groups more effectively during a foodborne outbreak investigation. Due to the high number of differences in modeling, further investigation will be needed to address the challenges of accessibility and affordability to different food options in South Carolina.

6.7 WORKS CITED

1. Knoops KT, de Groot LC, Kromhout D, Perrin AE, Moreiras-Varela O, Menotti A, van Staveren WA. Mediterranean diet, lifestyle factors, and 10-year mortality in elderly European men and women: the HALE project. JAMA. 2004;292:1433–1439.
2. Ruiz-Canela M, Zazpe I, Shivappa N, Hébert JR, Sánchez-Tainta A, Corella D, Salas-Salvadó J, Fitó M, Lamuela-Raventós RM, Fernández-Crehuet J, et al. Dietary inflammatory index and anthropometric measures of obesity in a population sample at high cardiovascular risk from the PREDIMED trial. Br J Nutr. 2015, 13, 984–995.
3. Park, YM; Steck SE, Fung TT, Zhang J, Hazlett LJ, Han K, Merchant AT. Mediterranean diet and mortality risk in metabolically healthy obese and metabolically unhealthy obese phenotypes. Int J Obes. 2016; Advance online publication. doi: 10.1038/ijo.2016.114
4. Trichopoulou A, Costacou T, Bamia C, Trichopoulos D. Adherence to a Mediterranean diet and survival in a Greek population. N Engl J Med. 2003;348:2599-2608.
5. Trust for America's Health and the Robert Wood Johnson Foundation. The State of Obesity in South Carolina. Published 2016. Accessed 24 Jun 2016. http://stateofobesity.org/states/sc/
6. CDC. State of South Carolina. Nutrition, Physical Activity, and Obesity Profile. Published 2012. Accessed 24 Jun 2016. http://www.cdc.gov/obesity/stateprograms/fundedstates/pdf/south-carolina-state-profile.pdf

7. CDC. Behavioral Risk Factor Surveillance System: Prevalence and Trend Data–Overweight and Obesity, U.S. Obesity Trends, Trends by State 2010. Accessed 24 Jun 2016. http://nccd.cdc.gov/NPAO_DTM/
8. CDC. MMWR September 2010 State–Specific Trends in Fruit and Vegetable Consumption Among Adults United States, 2000–2009. Accessed 24 Jun 2016. http://www.cdc.gov/mmwr/pdf/wk/mm5935.pdf
9. Yu D, Sonderman J, Buchowski MS, McLaughlin JK, Shu XO, Steinwandel M, Signorello LB, Zhang X, Hargreaves MK, Blot WJ, et al. Healthy eating and risks of total and cause-specific death among low-income populations of African-Americans and other adults in the southeastern United States: a prospective cohort study. PLoS Med 2015:12;e1001830.
10. Gao SK, Beresford SAA, Frank LL, Schreiner PJ, Burke GL, Fitzpatrick AL. Modifications to the Healthy Eating Index and its ability to predict obesity: the Multi-Ethnic Study of Atherosclerosis. Am J Clin Nutr. 2008;88:64–69.
11. Zamora D, Gordon-Larsen P, Jacobs DR, Popkin BM. Diet quality and weight gain among black and white young adults: the Coronary Artery Risk Development in Young Adults (CARDIA) Study (1985–2005). Am J Clin Nutr. 2010;92:784–793.
12. Bower KM, Thorpe RJ Jr., Rohde C, Gaskin DJ. The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the United States. Prev Med. 2014;58:33–39.
13. Darmon N, Drewnowski A. Does social class predict diet quality? Am J Clin Nutr. 2008;87:1107–1117.
14. Hiza HAB, Casavale KO, Guenther PM, Davis CA. Diet quality of Americans differs by age, sex, race/ethnicity, income, and education level. J Acad Nutr Diet. 2013;113:297–306.
15. Kirkpatrick SI, Dodd KW, Reedy J, Krebs-Smith SM. Income and race/ethnicity are associated with adherence to food-based dietary guidance among US adults and children. J Acad Nutr Diet. 2012;112:624–635.
16. Oldendick, B. (2013). *South Carolina Department of Health and Environmental Control Food Exposures Survey – Technical Report*. South Carolina: University of South Carolina's Institute for Public Service and Policy Research.
17. Nyachuba DG. Food-borne illness: is it on the rise? Nutr Rev. 2010; 68: 257-269.
18. Centers for Disease Control and Prevention (CDC). Surveillance for foodborne disease outbreaks—United States, 2007. *Morbidity and Mortality Weekly Report.* 2010;59(31):973–979.
19. Centers for Disease Control and Prevention. Foodborne Illness, Foodborne Disease, (sometimes called "Food Poisoning"). http://www.cdc.gov/foodsafety/facts.html#what. Published 2014. Accessed 1 April 2014.
20. Griffin AJ, McSorley SJ. Development of protective immunity to *Salmonella*, a mucosal pathogen with a systemic agenda. Mucosal Immunol. 2011;4:371–382.
21. Parry CM, Hien TT, Dougan G, White NJ, Farrar JJ. Typhoid fever. *N. Engl. J. Med.*. 2002; 347: 1770-1782.
22. US Department of Health and Human Services. Food Poisoning. http://www.foodsafety.gov/poisoning/. Published 2014. Accessed 1 April 2014.

23. Santos RL, Raffatellu M, Bevins CL, Adams LG, Tükel C, Tsolis RM, Bäumler AJ. Life in the inflamed intestine, *Salmonella* style. Trends Microbiol. 2009:17, 498–506.
24. U.S. Department of Agriculture Economic Research Service. South Carolina - Rural Definitions: State Level Maps. http://www.ers.usda.gov/datafiles/Rural_Definitions/StateLevel_Maps/SC.pdf. Published 2015. Accessed 15 Nov 2015.

6.8 TABLES

TABLE 6.1 STUDY POPULATION CHARACTERISTICS OF THE 875 RESPONDENTS - WEIGHTED*

Value	Total		Gender		Age		Hispanic		Race		Income		Urban/Rural	
	N (YES)	%	N (Male)	%	N (Under 45)	%	N (YES)	%	N (Caucasian)	%	N (<$50K)	%	N (Urban)	%
Poultry	816	93.26%	392	44.80%	463	52.91%	31	3.54%	516	58.97%	327	37.37%	618	70.63%
Meat	801	91.54%	402	45.94%	469	53.60%	31	3.54%	508	58.06%	328	37.49%	599	68.46%
Pork	654	74.74%	343	39.20%	387	44.23%	23	2.63%	411	46.97%	268	30.63%	481	54.97%
Seafood	364	41.60%	189	21.60%	193	22.06%	14	1.60%	202	23.09%	151	17.26%	266	30.40%
Eggs	783	89.51%	377	88.67%	457	53.35%	30	3.78%	492	56.16%	322	36.80%	588	67.21%
Dairy	853	97.49%	412	47.09%	489	55.89%	31	3.54%	541	61.83%	340	38.86%	641	73.26%
Fruit	834	95.31%	404	46.17%	474	54.17%	31	3.54%	525	60.00%	330	37.71%	624	71.31%
Vegetables	799	91.31%	379	43.31%	448	51.20%	31	3.54%	517	59.09%	323	36.91%	599	68.46%
Greens	715	81.71%	353	40.34%	404	46.17%	31	3.54%	459	52.46%	288	32.91%	537	61.37%
Snacks	748	85.49%	367	41.94%	452	51.66%	29	3.31%	484	55.31%	311	35.54%	555	63.43%
Nuts	748	85.49%	370	42.29%	438	50.06%	30	3.43%	501	57.26%	310	35.43%	557	63.66%
Grains	480	54.86%	227	25.94%	304	34.74%	17	1.94%	320	36.57%	204	23.31%	358	40.91%
Raw	827	94.51%	398	45.49%	465	53.14%	31	3.54%	523	59.77%	337	38.51%	623	71.20%
Frozen	792	90.51%	390	44.57%	457	52.23%	29	3.31%	503	57.49%	310	35.43%	597	68.23%
Pre-Pkg	858	98.06%	416	47.54%	494	56.46%	30	3.43%	541	61.83%	345	39.43%	640	73.14%

* Each response rounded to the nearest whole number

TABLE 6.2 RELATIVE RISKS OF EACH CATEGORY - WEIGHTED

Value	Gender			Age			Hispanic			Race			Income			Urban/Rural		
	RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits	
Poultry	1.025	0.9891	1.0628	1.001	0.9658	1.038	1.075	1.0557	1.0955	1.042	1.002	1.084	1.005	0.97	1.0422	1.083	1.0286	1.1406
Meat	0.94	0.9034	0.9789	1.075	1.0294	1.1223	1.096	1.0734	1.1192	1.048	1.003	1.096	1.041	1.001	1.0822	1.027	0.9774	1.0792
Pork	0.859	0.795	0.9279	1.101	1.0162	1.1927	0.979	0.7884	1.2154	1.022	0.944	1.108	1.043	0.966	1.1271	0.959	0.8808	1.0431
Seafood	0.872	0.7452	1.0208	0.866	0.7406	1.0132	1.082	0.7258	1.6116	0.754	0.646	0.881	1.07	0.913	1.2541	0.942	0.7901	1.1222
Eggs	1.018	0.9731	1.0658	1.068	1.0184	1.1208	1.0690	0.9865	1.1584	1.0211	0.9733	1.0712	1.047	1.002	1.0946	1.043	0.9855	1.1042
Dairy	1.011	0.9899	1.0333	1.023	0.9997	1.0463	1.013	0.9717	1.0562	1.052	1.024	1.081	0.994	0.972	1.0162	1.043	1.0089	1.0776
Fruit	1.01	0.9807	1.0403	1.003	0.9732	1.0328	1.051	1.0351	1.0672	1.028	0.996	1.062	0.98	0.95	1.0113	1.029	0.99	1.0688
Vegetables	1.049	1.0066	1.0935	0.974	0.9355	1.0144	1.099	1.0761	1.1228	1.109	1.057	1.165	1.02	0.979	1.0621	1.039	0.987	1.0945
Greens	0.972	0.9131	1.035	0.987	0.9268	1.0513	1.234	1.1945	1.2751	1.088	1.016	1.165	1.014	0.951	1.08	1.047	0.9698	1.1292
Snacks	0.983	0.9311	1.0384	1.162	1.0942	1.2332	1.086	0.9792	1.2045	1.11	1.044	1.181	1.067	1.012	1.1256	0.993	0.9336	1.0561
Nuts	0.969	0.9175	1.0232	1.077	1.0174	1.1408	1.125	1.039	1.2179	1.227	1.147	1.312	1.064	1.008	1.1221	1.006	0.9444	1.0714
Grains	1.054	0.9343	1.1893	1.314	1.1555	1.495	1.021	0.7421	1.4035	1.219	1.068	1.391	1.112	0.986	1.2538	1.016	0.8846	1.1677
Raw	1.017	0.9844	1.0501	0.98	0.9496	1.0117	1.061	1.0436	1.079	1.047	1.01	1.086	1.034	1.003	1.0664	1.058	1.0114	1.1071
Frozen	0.976	0.9348	1.0186	1.042	0.9964	1.0896	1.052	0.968	1.1439	1.054	1.005	1.105	0.963	0.92	1.0081	1.061	1.0024	1.1219
Pre-Pkg	1.007	0.988	1.0257	1.032	1.011	1.0542	0.994	0.9385	1.0525	1.033	1.009	1.057	1.006	0.988	1.0247	1.018	0.9919	1.044

TABLE 6.3 RELATIVE RISKS OF EACH CATEGORY WHEN CONTROLLING FOR URBAN VS. RURAL - URBAN

Value	Gender			Age			Hispanic			Race			Income		
	RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits	
Poultry	0.9884	0.9546	1.0233	1.0149	0.9789	1.0522	1.053	1.0342	1.072	0.983	0.9491	1.0171	0.9919	0.9563	1.0289
Meat	0.9507	0.9094	0.9939	1.0728	1.0218	1.1265	1.0877	1.0629	1.1131	1.053	0.9995	1.1092	1.0325	0.9882	1.0789
Pork	0.8552	0.7805	0.9369	1.0731	0.9762	1.1795	1.0252	0.7902	1.3301	1.029	0.9331	1.135	0.9935	0.904	1.0918
Seafood	0.8033	0.6674	0.967	0.7874	0.6555	0.946	0.53	0.2247	1.25	0.747	0.6224	0.8973	1.0099	0.8351	1.2213
Eggs	1.0199	0.9700	1.0724	1.0809	1.0241	1.1409	1.0350	0.9165	1.1688	1.0127	0.9597	1.0685	1.0382	0.9885	1.0903
Dairy	1.0063	0.9873	1.0256	1.0208	0.9995	1.0426	1.0155	1.0057	1.0254	1.032	1.0056	1.0589	0.9966	0.977	1.0167
Fruit	1.0024	0.9714	1.0344	1.023	0.9896	1.0575	1.0429	1.0262	1.0599	1.025	0.9889	1.0632	1.0118	0.9807	1.0438
Vegetables	1.0276	0.9822	1.075	0.9669	0.9254	1.0102	1.0874	1.0626	1.1128	1.084	1.0263	1.145	1.0134	0.9687	1.0602
Greens	0.9526	0.888	1.0219	0.985	0.9176	1.0573	1.2176	1.1741	1.2627	1.131	1.0408	1.2285	1.0321	0.9613	1.1081
Snacks	1.0132	0.9505	1.08	1.1523	1.0742	1.2362	1.1179	1.0036	1.2453	1.087	1.0101	1.1695	1.0377	0.9736	1.1059
Nuts	0.9588	0.9004	1.0209	1.0495	0.9829	1.1206	1.1144	1.0005	1.2412	1.217	1.1214	1.3199	1.0834	1.0194	1.1513
Grains	0.9867	0.8587	1.1339	1.2683	1.0932	1.4714	0.6904	0.3873	1.2309	1.226	1.0464	1.4362	1.2256	1.0688	1.4054
Raw	1.0029	0.971	1.0358	0.9742	0.9444	1.0049	1.0452	1.028	1.0627	1.03	0.9921	1.0696	1.0354	1.005	1.0668
Frozen	0.9581	0.9154	1.0027	1.0413	0.9921	1.0928	1.0703	1.0004	1.1451	1.035	0.9827	1.0894	0.9718	0.9251	1.0209
Pre-Pkg	1.0044	0.9854	1.0238	1.0175	0.9964	1.0389	1.0157	1.0058	1.0256	1.023	0.9985	1.0484	0.9979	0.9783	1.0179

TABLE 6.4 RELATIVE RISKS OF EACH CATEGORY WHEN CONTROLLING FOR URBAN VS. RURAL - RURAL

Value	Gender			Age			Hispanic			Race			Income		
	RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits		RR	Confidence Limits	
Poultry	1.1424	1.0353	1.2606	0.951	0.8637	1.0472	1.148	1.0901	1.2088	1.193	1.0756	1.3227	1.0662	0.9687	1.1736
Meat	0.9084	0.8304	0.9938	1.0777	0.9833	1.1811	1.1217	1.0703	1.1755	1.028	0.9404	1.1236	1.0714	0.9824	1.1684
Pork	0.8721	0.7549	1.0074	1.1866	1.0224	1.3772	0.8945	0.6124	1.3066	1.024	0.888	1.1815	1.1752	1.0211	1.3527
Seafood	1.0953	0.8127	1.4763	1.1308	0.8356	1.5303	1.9292	1.3948	2.6684	0.779	0.577	1.0509	1.2274	0.9116	1.6527
Eggs	1.0105	0.9123	1.1192	1.0274	0.9265	1.1393	1.1430	1.0445	1.2509	1.0273	0.9270	1.1386	1.0850	0.9812	1.1998
Dairy	1.0234	0.9609	1.09	1.0233	0.9595	1.0913	1.0252	0.9202	1.1421	1.093	1.0226	1.1679	0.9954	0.9341	1.0608
Fruit	1.0309	0.9614	1.1054	0.9422	0.8802	1.0086	1.0759	1.0368	1.1164	1.026	0.9561	1.101	0.9017	0.8359	0.9727
Vegetables	1.1132	1.0134	1.2228	0.9913	0.9031	1.0882	1.1356	1.0808	1.1933	1.167	1.058	1.2882	1.0475	0.9552	1.1486
Greens	1.0299	0.9	1.1785	0.9878	0.8632	1.1305	1.2859	1.1968	1.3817	0.969	0.8466	1.108	0.9725	0.8488	1.1143
Snacks	0.9029	0.8112	1.0049	1.1897	1.062	1.3327	1.0343	0.841	1.272	1.182	1.0578	1.3205	1.1517	1.0385	1.2771
Nuts	0.9985	0.8951	1.1138	1.1606	1.0338	1.3029	1.1456	1.0181	1.289	1.262	1.1221	1.4201	1.0129	0.9081	1.1299
Grains	1.2786	1.0022	1.6313	1.4551	1.1246	1.8826	1.5763	1.1846	2.0973	1.206	0.9442	1.5413	0.844	0.6595	1.0803
Raw	1.0554	0.9699	1.1486	0.9913	0.911	1.0786	1.1113	1.0626	1.1622	1.073	0.984	1.1698	1.0448	0.9612	1.1356
Frozen	1.0272	0.9268	1.1384	1.0369	0.9342	1.1509	1.0463	0.8648	1.2658	1.086	0.978	1.2063	0.9511	0.8563	1.0564
Pre-Pkg	1.012	0.9651	1.0612	1.0745	1.0196	1.1324	0.9655	0.8327	1.1194	1.052	1.0014	1.1059	1.0333	0.987	1.0818

TABLE 6.5 PARAMETERS SELECTED TO EACH FOOD MODEL

Outcome Variable	Exposure Variables Selected
Poultry	None
Meat	None
Pork	None
Seafood	Race
Eggs	None
Dairy	None
Fruit	None
Vegetables	None
Greens	None
Snacks	Age
Nuts	Race
Grains	Age, Race
Raw	None
Frozen	None
Pre-Pkg	None

TABLE 6.6 RELATIVE RISKS FOR EACH VARIABLE SELECTED TO THE SEAFOOD MODEL

	Forward, Backward, and Stepwise			
Parameter	***Relative Risk***	***Standard Error***	***Confidence Limits***	
Race	0.754	0.0795	0.6131	0.972

TABLE 6.7 RELATIVE RISKS FOR EACH VARIABLE SELECTED TO THE SNACKS MODEL

	Forward, Backward, and Stepwise			
Parameter	***Relative Risk***	***Standard Error***	***Confidence Limits***	
Age	1.1616	0.0868	1.0033	1.3449

TABLE 6.8 RELATIVE RISKS FOR EACH VARIABLE SELECTED TO THE NUTS MODEL

	Forward, Backward, and Stepwise			
Parameter	*Relative Risk*	*Standard Error*	*Confidence Limits*	
Race	1.2266	0.0953	1.0533	1.4283

TABLE 6.9 RELATIVE RISKS FOR EACH VARIABLE SELECTED TO THE NUTS MODEL

	Forward, Backward, and Stepwise			
Parameter	***Relative Risk***	***Standard Error***	***Confidence Limits***	
Age	1.3261	0.1257	1.1012	1.5968
Race	1.2342	0.1197	1.0204	1.4926

CHAPTER 7

CONCLUSION

7.1 CONCLUSION

While foodborne pathogens and the study of outbreaks themselves are not innovative, the ways they are being investigated continue to evolve. Random Forests is a relatively new data driven machine-learning tool to identify predictive patterns in big data that is used in many diverse fields. Its application to foodborne disease outbreak investigations can potentially help to identify foods causing illness quickly. We have applied Random Forests to identify food and environmental exposures associated with *Salmonella* outbreaks in South Carolina using outbreak investigation data collected by DHEC (Aim 1). Logistic regression is commonly used in epidemiological studies as it can provide relative risk and odds ratios, but it can be limited when analyzing datasets like the one in this study because of its high number of exposures and the multiple interactions they can have. It would have been nearly impossible to add all relevant interactions to a logistic regression model considering there were 207 relevant exposures. Random forests is a popular method used in biomedical studies, but its use in epidemiological studies is minimal.[1-4] The random forests method was used to analyze the complex relationships in food outbreak data, handle a high volume of exposures, and deal with missing data.

We used a random forests algorithm to find exposures for cases in three *Salmonella* outbreaks that occurred in South Carolina in the past 5 years. The random forests algorithm generated lists of the top 30 suspected exposures out of 207 individual food and environmental exposures that contained the foods associated with each of the three outbreaks that were analyzed. Random forests may aid in investigations of foodborne outbreaks and aid in quicker identification of their causes.

In addition, we studied dietary patterns derived from questionnaires that DHEC administered to different groups of people in South Carolina. The first dietary pattern was derived from a food exposure questionnaire that had been administered to individuals as part of outbreak investigations. Food groups associated with *Salmonella* outbreaks were identified, and then described by demographic and other characteristics (Aim 2).

The aim of this study was to evaluate the food causes of *Salmonella* outbreaks in South Carolinians using two standard definitions of an outbreak. The FDA defines an outbreak as two or more cases of foodborne illness caused by the same organism that occur within a limited period of time and are associated with either the same food or same food service operation.[5] This study interpreted this as two or more cases of the same organism that occurred within 30 days and used this to find the food commonalities. Since DHEC has more knowledge of sporadically occurring cases, they are more stringent with their outbreak definition and limit it by serotype, date, and foods eaten. This definition was provided to us in the data collected from DHEC.

Using two separate definitions yielded only dairy as a food common between both. In the DHEC definition, dairy was only 0.5 times as likely to cause an outbreak whereas in the FDA definition, that number jumped to 1.14 times as likely. This does tell us that unpasteurized milk and other dairy products can be a food that is associated with outbreaks in the state. Meat and dairy are also associated when using the DHEC definition. The differences in the relative risks in the two definitions are also letting us know that living in an urban or rural environment and being Hispanic can also be associated with foodborne illness.

The second food group pattern was derived from a questionnaire that was similar to the one DHEC had used for the outbreak investigation, but was administered to a representative sample of individuals living in South Carolina. We then described the demographic, social and geographic predictors of consumption of these food groups by South Carolina residents (Aim 3). The foods eaten by South Carolinians varied widely when participants were placed into demographic groups. By using a representative sample that contains geographic and socio-demographic diversity, we can use the information to potentially affect change in the food and dietary programs available to South Carolinians.

We aimed to look at the demographic, social, and geographic determinants of dietary patterns in the South Carolina population. Participants that identified as Hispanic and Caucasians consumed the highest amount of fresh food while those under 45 years of age and Caucasians consumed the most pre-packaged food. Caucasians and those that live in an urban environment eat the most frozen food relative to other races and those in a rural environment, respectively. When controlling for an urban or rural setting, the results changed to show that while the fresh food demographics did not change, the frozen and pre-packaged relative risks did. It was those who identified as Hispanic that were more likely to consume frozen or pre-packaged items relative to those that did not identify as Hispanic. The rural setting more closely resembled the state as a whole. Seafood, snacks, nuts, and grains were associated with the demographic characteristics of age and race. The differences in relative risks of foods consumed when controlling for demographic characteristics is telling us that many different characteristics can be associated with access to different types of food.

In summary, we have used Random Forests to analyze data that are routinely collected during foodborne outbreak investigations. This new application of Random Forests can make identification of foods responsible outbreaks more efficient. Also, the information characterizing food exposure data collected by DHEC as part of its surveillance, will help in interpreting data collected in outbreak investigations. This information will address the challenges of a rural southern state with a high obesity rate by using a representative sample that contains geographic and socio-demographic diversity and using said information to help affect change in the programs available. To the best of our knowledge, no other study has attempted to collect this information and make this comparison. The results of this study can potentially improve foodborne disease outbreak investigations in South Carolina.

7.2 WORKS CITED

1. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7:3.
2. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics.* 2007; 23(19): 2507-2517.
3. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*. 2007; 8: 25.
4. Boulesteix AL, Bender A, Bermejo JL, Strobl C. Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics.* 2012; 13 (3): 292-304.
5. U.S. Food and Drug Administration. What You Should Know About Government Response to Foodborne Illness Outbreaks. http://www.fda.gov/Food/ResourcesForYou/Consumers/ucm180323.htm. Published 2014. Accessed 20 October 2014.

APPENDIX A – PROTECTION OF HUMAN SUBJECTS

Target Population

The target population for this research was individuals aged two and older who may have been a part of a foodborne outbreak in South Carolina and had both a specimen analyzed by DHEC BoL and filled out a questionnaire administered by DHEC. If these conditions were not met, the person was not be included in the analysis. To be considered part of an outbreak, two or more cases of foodborne illness had to have occurred during a limited period of time with the same organism and that were associated with either the same food service operation, such as a restaurant, or the same food product.[57]

Recruiting Plans

Recruitment of individuals for this research was not necessary because the investigators only used data that is already in existence. Only de-identified data was collected to prevent from exposing any private information.

Existing Data/Samples

Investigators used existing data from DHEC's Division of Acute Disease Epidemiology (DADE). While this data contains direct identifiers, none of that information was collected. The data was de-identified and sent over from DADE as an excel workbook and contained a unique key for each individual for easy tieback to the original data located in DADE.

Consent/Assent

For this research, consent from persons who have participated in this study was not necessary since the data is already in existence. Only de-identified data was collected to prevent from exposing any person's private information.

Potential Risks

We believe that there is minimal risk to study participants, because information that will be collected will not contain any personal information and the investigators using the data will not have access to any of the personal information that may be on file at DHEC. The proposed research falls under an exemption category as nonhuman subjects research (AIM 1, AIM 2, and AIM 3). An application has already been submitted and approved by both the University of South Carolina, Institutional Review Board (IRB) and the DHEC IRB. The human subjects data used for AIMS 1, 2, and 3 of this proposal are part of the data previously collected by DHEC's Division of Acute Disease Epidemiology (DADE). DADE will protect the individual study participant's identifying data by assigning a unique key to each person before sending the data over to investigators. No individual will be identified in any publications resulting from this study.

Potential Benefits

There was no direct benefit to participants as a result of this study. However, the information obtained from this study added to the body of scientific knowledge about this important area of research. This research also provided information that can be used to inform foodborne outbreak investigators of the possible food cause of an outbreak much quicker and more effectively than previous methods.

Confidentiality

As mandated by the state and the Institutional Review Board (IRB) of the University of South Carolina as well as the IRB at the South Carolina Department of Environmental Health and Control (DHEC), authorized persons trained in the Health Insurance Portability and Accountability Act (HIPAA) confidentiality procedures retrieved the data. Study investigators were required to abide by the guidelines set forth by the state and school with regards to the security and confidentiality of the data. The principle investigator was the only person collecting the data. Unique identifiers were created for each patient at DHEC. The data sent from DHEC was kept on a secure drive and only accessed on password-protected computers. The data was not stored on a local hard drive or network drive. All data analysis by the investigators was conducted at University of South Carolina School of Public Health using password-protected computers. Any data that may have been printed out was shredded immediately upon completion of its use.

APPENDIX B – DHEC HYPOTHESIS-GENERATING QUESTIONNAIRE FOR CHAPTERS 4 AND 5

D H E C
PROMOTE PROTECT PROSPER
South Carolina Department of Health and Environmental Control

DHEC Hypothesis Generating Questionnaire

○ Salmonella ○ STEC ______ Enter serotype if known (Enteritidis, Typhimurium, O157:H7, etc.

DHEC Region ______ SC PFGE Pattern (if known): ______

Section 1: Interviewer information *(Questions 1-5 to be completed by interviewer prior to questionnaire administration)*

1. PulseNet ID # (if known): ______ 2. CHESS Investigation ID #: ______

3. Date of Interview *(must enter MM/DD/YYYY)* ______

4. Interviewer Information Name ______ Phone Number ______

5. Before this interview how many times has the case been interviewed about their illness? ○ None ○ Once ○ Twice ○ Three Times ○ Other (Specify) ______ ○ Unknown

6. Respondent was: ○ Self ○ Parent ○ Spouse ○ Other (Specify) ______

Section 2: Demographic Data: I'd like to begin by asking a few questions about yourself (your child) and your household.

1.Last Name: ______ 2. First Name: ______

3.Street Address ______

4. City ______ 5. State SC 6. County ______ 7. Zip Code ______

8. Phone Number ______ ○ Home ○ Work ○ Cell

9. Sex ○ Male ○ Female ○ Unknown 10. Age ______

11. Birth month *(must enter 1-12)* ______ 12. Birth day *(must enter 1-31)* ______ 13. Birth year *(must enter YYYY)* ______

14. Hispanic/Latino origin ○ Yes ○ No ○ Unknown

15. How do you describe your race? ○ White ○ Black/African American ○ American Indian/Alaska Native ○ Asian ○ Native Hawaiian/Other Pacific Islander ○ Other ○ Unknown

16. Occupation ______

YES	NO	Don't Know	High-Risk Exposures
○	○	○	17. Are you a food handler (workplace, volunteer, farm work, etc.)?
		If YES, Where?	
○	○	○	18. Are you associated with a childcare center or school?
		If YES, Where?	

19. Primary source of drinking water? ○ Municipal ○ Bottled ○ Private Well ○ Other ○ Unknown

20. If case is an infant (Check all that apply)
☐ Breastfed ☐ Formula Specify brand: ______
☐ Jarred baby food Specify brand/ varieties: ______
☐ Cow's milk ☐ Other Specify: ______

DHEC 0928 (03/2011) Page 1 of 13

Section 3: Clinical Information: Now I have a few questions about your (your child's) illness.

1. What date did you first feel sick? (any symptoms) *(must enter MM/DD/YYYY)*

YES	NO	Don't Know	Did/Were you (your child)
◯	◯	◯	2. Have any diarrhea (defined as at least 3 loose stools in 24 hours)
			2a. What day did it start? *(must enter MM/DD/YYYY)*
◯	◯	◯	2b. Have any blood in the stool?
◯	◯	◯	3. Have any vomiting?
◯	◯	◯	4. Have a fever? 4a. Highest recorded temperature
◯	◯	◯	5. Hospitalized overnight?
			5a. If YES, Where? 5b. Admitted *(MM/DD/YYYY)* 5c. Discharged *(MM/DD/YYYY)* 6 . How many days total were you sick? or ☐ Still Ill
◯	◯	◯	7. Have any close contact with anyone with diarrhea or vomiting? 7a. When was this person ill ◯ less than 24 hours before you ◯ ≥ 24 hours before you ◯ Unknown 7b. Name/Contact Info

Section 4: General Exposure Information

YES	NO	Don't Know	In the 7 days before your illness, did you (your child)
◯	◯	◯	1. Attend any large gatherings? (weddings, parties, etc.)
			Event 1: Date *(MM/DD/YYYY)* Event 2: Date *(MM/DD/YYYY)* Event 3: Date *(MM/DD/YYYY)*
◯	◯	◯	2. Spend time travelling away from your home?
			2a. List all SC cities where you travelled. Enter SC cities:
			2b. List all US states where you travelled. ◯ Did not travel outside SC ◯ Unknown List US States (2-ltr abrvs) & dates
			2c. List all countries outside the US where you travelled. ◯ Did not travel outside US ◯ Unknown List countries & travel dates
◯	◯	◯	3. Have any contact with animals (pets, wild animals, farm animals, reptiles)?
			Type of Animal Brand Pet Food/Treats
◯	◯	◯	4. Swim or wade (rivers, lakes, swimming pools, etc.)?
			Specify
◯	◯	◯	5. Consume raw dairy products (unpasteurized)?
			Product Type/Brand Purchase Place

If case spent the entire 7 days before illness outside of the US, please note the countries visited and skip to end of questionnaire.
If case spent part of the 7 days before illness onset outside the US, please collect food information only on foods eaten in the US.

Section 5: Food allergies, special diets, vitamins, & supplements: Now I have a few questions about food allergies and any special diets you (your child) may follow. I will also ask a few questions about vitamins and supplements you may have taken in the 7 days before your illness began.

YES	Maybe	NO	Don't Know	Did you (your child) have:
○	○	○	○	1. Any allergies that prevent you from eating a certain food(s)?
				1a. What foods? *Please check all that apply* ☐ milk ☐ eggs ☐ peanuts ☐ tree nuts ☐ fish ☐ soy ☐ wheat ☐ shellfish ☐ other (Specify)
○	○	○	○	2. Vegetarian or vegan diet?
○	○	○	○	3. Special or restricted diet (medical, weight-loss, religious, cultural, etc.)?
				3a. Please describe: ☐ Unknown
○	○	○	○	4. Any commercially bottled water in personal-sized containers?
○	○	○	○	5. Any commercially bottled water in large, multi-user tanks or water coolers?
○	○	○	○	6. Any vitamins, nutritional or herbal supplements, such as teas, tablets, and pills, etc.? 6a. Please describe type, variety, brand: ☐ Unknown

Section 5 Comments. *Please fill in any comments/notes from this section in the space provided*

For Sections 6 and 7: Read each type of store, point of purchase, or food outlet in the top section and ask respondent to list names for each category. The lists of store/restaurant types are meant to prompt the respondent. Please list the names of all points of purchase/restaurants mentioned, regardless of category, in the space provided below. ***You do not need to record a yes or no response for each category, only record the specific names and approximate locations reported in the space below.***

Section 6: Sources of food at home: Now I have a few questions about where the food came from that you ate **at home** in the 7 days before your illness began. This isn't necessarily where you shopped during that week, but where what you actually ate came from. I'm going to list several types of stores, for each type please tell me the names of each store you would have eaten food from during the 7 days before you were sick.

1. Did you (your child) eat foods from?

YES	Maybe	NO	Don't Know	
○	○	○	○	Grocery Stores or Supermarkets
○	○	○	○	Warehouse stores such as Costso or Sam's Club
○	○	○	○	Small markets or mini markets (gas stations, etc.)
○	○	○	○	Ethnic specialty markets (Indian, Mexican, Asian, etc.)
○	○	○	○	Health food stores or Co-ops
○	○	○	○	Fish or meat specialty shops (butcher's shop, etc.)
○	○	○	○	Farmer's markets, roadside markets, food purchased directly from farm
○	○	○	○	Any other sources of food at home in the 7 days before illness

If YES to any of the above, enter name and address of store/food sources at home below.

Name of store/food source 1: Location:

Name of store/food source 2: Location:

Name of store/food source 3: Location:

Name of store/food source 4: Location:

Name of store/food source 5: Location:

Section 7: Sources of food outside the home: Now I have a few questions about where the food came from that you ate **outside your home** such as restaurants or fast food chains. I'm going to list several types of restaurant, for each type please tell me the names of each place you would have eaten food from during the 7 days before you were sick.

1. Did you (your child) eat foods from?

YES	Maybe	NO	Don't Know	
○	○	○	○	National fast food chain
○	○	○	○	Mexican-style
○	○	○	○	Italian
○	○	○	○	Seafood
○	○	○	○	Chinese, Japanese, Indian, other Asian-style
○	○	○	○	BBQ or Homestyle
○	○	○	○	Steakhouse or Grill
○	○	○	○	Breakfast or Pancake House
○	○	○	○	All-you-can-eat Buffet
○	○	○	○	School, hospital or any institutional setting
○	○	○	○	Sandwich shop or deli
○	○	○	○	Any event that was catered
○	○	○	○	Any other food away from home

If YES to any of the above, enter name and address of restaurants/food source away from home below.

Name of restaurant 1: ________ Location: ________

Name of restaurant 2: ________ Location: ________

Name of restaurant 3: ________ Location: ________

Name of restaurant 4: ________ Location: ________

Name of restaurant 5: ________ Location: ________

Name of restaurant 6: ________ Location: ________

Name of restaurant 7: ________ Location: ________

Name of restaurant 8: ________ Location: ________

Name of restaurant 9: ________ Location: ________

Name of restaurant 10: ________ Location: ________

Section 6 & 7 Comments ________

OSS Sentinel Site Participants Continue to end of form. Non-OSS sites may stop here for Salmonella.

Section 8: Meat and Poultry: Now I have a few questions about meat and poultry that you (your child) might have eaten in the 7 days before your (your child's) illness began. This does not include canned items, but the meat and poultry could have been fresh, frozen, or could have been eaten as part of dish. You (your child) could have eaten these either in your home or outside the home. As I read each food, please answer as yes, no, may have eaten, or can't remember eating the food in the 7 days before you (your child) got sick.

First, I have questions about CHICKEN & OTHER POULTRY products.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
○	○	○	○	1. Whole chicken or cut chicken pieces/parts?
				1a. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Not applicable (did not eat at home)
				1b. If eaten outside the home, where? List name(s) and location(s): ___ ☐ Not applicable (did not eat outside the home)
○	○	○	○	2. Ground chicken?
○	○	○	○	3. Breaded chicken products, such as chicken tenders, strips, or nuggets?
○	○	○	○	4. Stuffed, frozen chicken products, such as chicken Kiev or chicken Cordon Bleu?
○	○	○	○	5. Any other frozen chicken products?
○	○	○	○	6. Duck, game hen, or squab?
○	○	○	○	7. Whole turkey or cut turkey pieces/parts?
				7a. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Not applicable (did not eat at home)
				7b. If eaten outside the home, where? List name(s) and location(s): ___ ☐ Not applicable (did not eat outside the home)
○	○	○	○	8. Ground Turkey?

Section 8 Chicken/Poultry Comments. *Please fill in comments/notes from this section in the space provided* ___

Now I have questions about BEEF products

YES	Maybe	NO	Don't Know	In the 7 days before the illness began, did you (your child) eat any:
○	○	○	○	9. Beef steaks or roasts?
				9a. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Purchased Frozen ☐ Purchased Fresh ☐ Was pink or red inside when eaten ☐ Not applicable (did not eat at home)
				9b. If eaten outside the home, where? List name(s) and location(s): ___ ☐ Not applicable (did not eat outside the home)
○	○	○	○	10. Pre-made or pre-formed hamburger patties at home?
				10a. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Was pink or red inside when eaten
○	○	○	○	11. Any fresh hamburger patties at home?
				11a. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Was pink or red inside when eaten

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
○	○	○	○	12. Any dish with ground beef at home, such as casseroles, tacos, soups, or pasta sauces?
				12a. If eaten at home: Please describe the dish: Place purchased from (names, locations): ☐ Not applicable (did not eat at home)
○	○	○	○	13. Any ground beef outside the home? This could include foods such as hamburger or other dishes such as casseroles, tacos, soups, or pasta sauces.
				13a. Where did you eat this? Place purchased from (names, locations): ☐ Not applicable (did not eat at home)
Section 8 Beef Comments. *Please fill in comments/notes from this section in the space provided*				

Now I have questions about PORK, LAMB, & OTHER MEAT products

YES	Maybe	NO	Don't Know	In the 7 days before the illness began, did you (your child) eat any:
○	○	○	○	14. Ground pork?
○	○	○	○	15. Any other pork product?
				15a. What was the type, variety, brand: ☐ Unknown
○	○	○	○	16. Lamb?
○	○	○	○	17. Bacon?
○	○	○	○	18. Sausage?
○	○	○	○	19. Hot dogs, corn dogs, polish sausage, kielbasa, or similar product?
○	○	○	○	20. Pepperoni? This could have been on a sandwich or pizza.
○	○	○	○	21. Any other Italian-style meats, such as salami or prosciutto?
○	○	○	○	22. Bologna, pastrami, or corned beef?
○	○	○	○	23. Store-bought, dried meat strips or jerky?
○	○	○	○	24. Pre-packaged deli meats?
				24a. If eaten at home, what was the: Type, variety, brand: Place purchased from (names, locations): ☐ Not applicable (did not eat at home)
○	○	○	○	25. Any other deli-sliced meat (not pre-packaged)?
				25a. If eaten at home, what was the: Type, variety, brand: Place purchased from (names, locations): ☐ Not applicable (did not eat at home)
				25b. If eaten outside the home, where? List name(s) and location(s): ☐ Not applicable (did not eat outside the home)
○	○	○	○	26. Any other meat and/or poultry products, not mentioned already?
Section 8 Pork, Lamb, Other Meats Comments. *Please fill in comments/notes from this section in the space provided*				

Section 9: Fish and Seafood: Now I have some questions about fish and seafood you (your child) might have eaten in the 7 days before your (your child's) illness began. You (your child) may have eaten this at home or away from home. This does not include canned items, but these foods could have been eaten alone or as part of a dish, sauce, or dip. As I read each food, please answer as yes, no, may have eaten, or can't remember eating the food in the 7 days before you (your child) got sick.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
○	○	○	○	1. Fresh or fresh-frozen fish?

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
◯	◯	◯	◯	2. Smoked or dried fish?
◯	◯	◯	◯	3. Shrimp or prawns?
◯	◯	◯	◯	4. Crab, lobster, or crayfish?
◯	◯	◯	◯	5. Oysters?
◯	◯	◯	◯	6. Clams, mussels, scallops, or other shellfish?
◯	◯	◯	◯	7. Sushi (with raw fish or seafood)?
◯	◯	◯	◯	8. Frozen fish product (fish sticks, nuggets, etc.)?
◯	◯	◯	◯	9. Any other fish or seafood?
				9a. What was the type, variety, brand: ☐ Unknown

Section 9 Comments. *Please fill in comments/notes from this section in the space provided*

Section 10: <u>Eggs, Dairy, and Cheese</u>: Now I have a few questions about eggs, dairy, and cheese products you (your child) might have eaten in the 7 days before your (your child's) illness began. You (your child) could have eaten these either in your home or away from home. As I read each food, please answer as yes, no, may have eaten, or can't remember eating the food in the 7 days before you (your child) got sick.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
◯	◯	◯	◯	1. Eggs or egg-containing dishes?
◯	◯	◯	◯	1a. Were they raw, "runny", or "over-easy"?
				1b. If eaten <u>at home</u>, what was the: Type, variety, brand: Place purchased from (names, locations): ☐ Not applicable (did not eat <u>at home</u>)
				1c. If eaten <u>outside the home</u>, where? List name(s) and location(s): Describe the dish: ☐ Not applicable (did not eat <u>outside the home</u>)
◯	◯	◯	◯	2. Anything made with raw eggs (cookie dough, cake batter, sauces, homemade ice cream or mayo, etc.)?
◯	◯	◯	◯	3. Milk?
				3a. What was the type, variety, brand: ☐ Unknown
				3b. Raw or unpasteurized? ◯ Yes ◯ No ◯ Maybe ◯ Unknown
◯	◯	◯	◯	4. Ice cream or ice cream products?
				4a. If eaten <u>at home</u>, what was the: Type or brand (bar, tub, carton, etc.) Variety or flavor? ☐ Unknown
◯	◯	◯	◯	5. Frozen yogurt?
◯	◯	◯	◯	6. Yogurt drinks?
◯	◯	◯	◯	7. Any other yogurt?
◯	◯	◯	◯	8. Prepackaged, shredded cheese?
				8a. What was the type, variety, brand: ☐ Unknown
◯	◯	◯	◯	9. Processed, sliced cheese?
◯	◯	◯	◯	10. Block-type cheese (cheddar, Swiss, Colby, Jack, etc.)?
				10a. What was the type, variety, brand: ☐ Unknown
◯	◯	◯	◯	11. String-type cheese?
◯	◯	◯	◯	12. Cottage cheese?
◯	◯	◯	◯	13. Cheese curds?
◯	◯	◯	◯	14. Feta cheese? This could have been part of a dish or salad.

YES	Maybe	NO	Don't Know	In the 7 days before the illness began, did you (your child) eat any:
◯	◯	◯	◯	15. Blue veined cheese (gorgonzola, bleu)?
◯	◯	◯	◯	16. Fresh or dried Parmesan, Romano, or similar cheese?
◯	◯	◯	◯	17. Cheese from raw/unpasteurized milk (homemade, farm-fresh, door-to-door cheeses)?
◯	◯	◯	◯	18. Mexican-style soft cheese (queso fresco, queso blanco)?
◯	◯	◯	◯	18b. Was it homemade?
◯	◯	◯	◯	19. Any other gourmet or artisanal cheese?
				19a. What was the type, variety, brand: ☐ Unknown
◯	◯	◯	◯	20. Any other dairy product?
				20a. What was the type, variety, brand: ☐ Unknown

Section 10 Comments. *Please fill in comments/notes from this section in the space provided*

Section 11: Fresh fruits: Now I have some questions about fresh fruits, not canned, cooked, or frozen, you (your child) might have eaten in the 7 days before your (your child's) illness began. You (your child) could have eaten these either in your home or away from home. As I read each food, please answer as yes, no, may have eaten, or can't remember eating the food in the 7 days before you (your child) got sick.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
◯	◯	◯	◯	1. Apples?
				1a. What was the type, variety, brand: ☐ Unknown
◯	◯	◯	◯	2. Grapes?
				2a. What was the type, variety, brand: ☐ Unknown
◯	◯	◯	◯	3. Pears?
◯	◯	◯	◯	4. Peaches?
◯	◯	◯	◯	5. Nectarines?
◯	◯	◯	◯	6. Apricots?
◯	◯	◯	◯	7. Plums?
◯	◯	◯	◯	8. Oranges?
◯	◯	◯	◯	9. Grapefruit?
◯	◯	◯	◯	10. Tangerines?
◯	◯	◯	◯	11. Fresh lemon or lime? This could include garnishes on a drink, etc.
◯	◯	◯	◯	12. Strawberries?
◯	◯	◯	◯	13. Raspberries?
◯	◯	◯	◯	14. Blueberries?
◯	◯	◯	◯	15. Blackberries?
◯	◯	◯	◯	16. Cherries?
◯	◯	◯	◯	17. Any other fresh berries?
				17a. What was the type, variety, brand: ☐ Unknown
◯	◯	◯	◯	18. Cantaloupe?
◯	◯	◯	◯	19. Honeydew melon?
◯	◯	◯	◯	20. Watermelon?
◯	◯	◯	◯	21. Precut melon or melon salad? Sometimes served on salad bars or breakfast buffets
◯	◯	◯	◯	22. Any other melon?

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
○	○	○	○	23. Pineapple?
○	○	○	○	24. Mango?
○	○	○	○	25. Coconut (whole or shredded)?
○	○	○	○	26. Any other tropical fruit (kiwi, papaya, guava, pomegranate, etc.)?
				26a. What was the type, variety, brand: ☐ Unknown
○	○	○	○	27. Frozen berries?
○	○	○	○	28. Other frozen fruit?
○	○	○	○	29. Raisins?
○	○	○	○	30. Other dried fruit?
				30a. What was the type, variety, brand: ☐ Unknown
○	○	○	○	31. Apple juice (not from concentrate)?
○	○	○	○	32. Orange juice (not from concentrate)?
○	○	○	○	33. Any other juice (not from concentrate)?
				33a. What was the type, variety, brand: ☐ Unknown
○	○	○	○	34. Juice from frozen concentrate?
○	○	○	○	35. Any unpasteurized or raw juices or ciders?

Section 11 Comments. *Please fill in comments/notes from this section in the space provided*

Section 12: Fresh Vegetable: Now I have some questions about fresh vegetables you (your child) might have eaten raw or uncooked in the 7 days before your (your child's) illness began. You (your child) could have eaten these either in your home or away from home. This does not include canned items, but these foods could have been eaten alone or as part of a dish. I am only interested in vegetables that are not grown at home. As I read each food, please answer as yes, no, may have eaten, or can't remember eating the food in the 7 day before you (your child) got sick.

First, I have questions about TOMATOES & LEAFY GREENS

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
○	○	○	○	1. Fresh tomatoes?
				1a. If eaten at home, what was the type, variety: ☐ Red Round ☐ Roma ☐ Cherry ☐ Grape ☐ 'Vine-ripe', sold on the vine ☐ other (Specify) ☐ Unknown Place purchased from (names, locations): ☐ Not applicable (did not eat at home)
				1b. If eaten outside the home, where? List name(s) and location(s): ☐ Not applicable (did not eat outside the home)
○	○	○	○	2. Fresh tomatoes on sandwich, burger, or salad?
○	○	○	○	3. Fresh salsa or pico de gallo (not from a jar or can)?
				3a. If eaten at home, what was the: Type, variety (red, green): Place purchased from (names, locations): ☐ Not applicable (did not eat at home)
				3b. If eaten outside the home, where? List name(s) and location(s): ☐ Not applicable (did not eat at home)
○	○	○	○	4. Guacamole?
○	○	○	○	5. Fresh, uncooked leafy greens such as spinach, lettuce, etc.?
				5a. Prepackaged or loose? ☐ Prepackaged ☐ Loose ☐ Unknown
				5b. If eaten at home, what was the: Type, variety, brand: Place purchased from (names, locations): ☐ Not applicable (did not eat at home)
				5c. If eaten outside the home, where? List name(s) and location(s): ☐ Not applicable (did not eat outside the home)

○	○	○	○	6. Lettuce on a sandwich, burger, or as garnish?
○	○	○	○	7. Iceberg lettuce?
				7a. Prepackaged or head/loose? ☐ Prepackaged ☐ Head/Loose ☐ Unknown
				7b. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Not applicable (did not eat at home)
				7c. If eaten outside the home, where? List name(s) and location(s): ___ ☐ Not applicable (did not eat outside the home)
○	○	○	○	8. Romaine lettuce?
				8a. Prepackaged or loose? ☐ Prepackaged ☐ Loose ☐ Unknown
				8b. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Not applicable (did not eat at home)
				8c. If eaten outside the home, where? List name(s) and location(s): ___ ☐ Not applicable (did not eat outside the home)
○	○	○	○	9. Fresh spinach?
				9a. Prepackaged or loose? ☐ Prepackaged ☐ Loose ☐ Unknown
				9b. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Not applicable (did not eat at home)
				9c. If eaten outside the home, where? List name(s) and location(s): ___ ☐ Not applicable (did not eat outside the home)
○	○	○	○	10. Cabbage?
○	○	○	○	11. Other leafy lettuce (red, butter, radicchio, "spring mix", "baby" salad greens, etc.?

Section 12 Tomato & Leafy Greens Comments. *Please fill in comments/notes from this section in the space provided*

Now I have questions about herbs & sprouts you (your child) may have eaten in the 7 days before illness began. Remember, these could have been part of a dish such as pesto, salsa, sauces, etc. We are interested in fresh herbs, not dried or bottled herbs.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
○	○	○	○	12. Fresh Basil?
○	○	○	○	13. Fresh cilantro?
○	○	○	○	14. Other fresh herbs (parsley, sage, thyme, dill, etc.)?
				14a. What was the type, variety, brand: ___ ☐ Unknown
○	○	○	○	15. Alfalfa sprouts?
				15a. If eaten at home, what was the: Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Not applicable (did not eat at home)
				15b. If eaten outside the home, where? List name(s) and location(s): ___ ☐ Not applicable (did not eat outside the home)
○	○	○	○	16. Other sprouts (bean, clover, broccoli, daikon radish, etc.)?
				16a. If eaten at home, what was the:Type, variety, brand: ___ Place purchased from (names, locations): ___ ☐ Not applicable (did not eat at home)
				16b. If eaten outside the home, where? List name(s) and location(s): ___ ☐ Not applicable (did not eat outside the home)

Section 12 Herbs & Sprouts Comments. *Please fill in comments/ notes from this section in the space provided*

Next I have a few questions about other fresh vegetables you (your child) may have eaten in the 7 days before your illness began.

YES	Maybe	NO	Don't Know	In the 7 days before the illness began, did you (your child) eat any:
◯	◯	◯	◯	17. Cucumbers, zucchini, squash?
◯	◯	◯	◯	18. Bell peppers (green, red, orange, or yellow)?
◯	◯	◯	◯	19. Hot chili/chile peppers (such as jalapenos or serranos)?
◯	◯	◯	◯	20. Celery?
◯	◯	◯	◯	21. "Mini" carrots? These are often peeled and sold in a sealed bag.
◯	◯	◯	◯	22. Other fresh carrots?
◯	◯	◯	◯	23. Other raw root vegetables (radishes, beets, turnips, etc.)?
				23a. What was the type, variety, brand: ☐ Unknown
◯	◯	◯	◯	24. Fresh, raw peas? May be shelled or in the pod
◯	◯	◯	◯	25. Broccoli?
◯	◯	◯	◯	26. Cauliflower?
◯	◯	◯	◯	27. Raw onions (white, yellow, or red/purple)?
◯	◯	◯	◯	28. Raw green onions/scallions?
◯	◯	◯	◯	29. Fresh or dried mushrooms?

Section 12 Other Vegetable Comments. *Please fill in comments/notes from this section in the space provided*

Section 13: Frozen Foods: Now I have a few questions about frozen foods you (your child) might have eaten in the 7 days before your (your child's) illness began. You (your child) could have eaten these either in your home or outside the home. As I read each food, please answer as yes, no, may have eaten, or can't remember eating the food in the 7 days before you (your child) got sick.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
◯	◯	◯	◯	1. Frozen vegetables (in bag or box)?
◯	◯	◯	◯	2. Frozen pot pies?
◯	◯	◯	◯	3. Frozen pizza?
◯	◯	◯	◯	4. Frozen Mexican-style foods (burritos, etc.)?
◯	◯	◯	◯	5. Frozen snack foods like mozzarella sticks, jalapeno poppers, potato skins, or hot pockets?
◯	◯	◯	◯	6. Frozen breakfast items (waffles, breakfast sandwiches, etc.)?
◯	◯	◯	◯	7. Frozen vegetarian foods such as a garden burger?
◯	◯	◯	◯	8. Frozen pre-mixed meals in a bag or box (stir fry, pasta meals, etc.)?
◯	◯	◯	◯	9. Frozen dinners or box entrees?
◯	◯	◯	◯	10. Other frozen, prepackaged product not mentioned previously?
				10a. What was the type, variety, brand: ☐ Unknown

Section 13 Comments. *Please fill in comments/notes from this section in the space provided*

Section 14: Nuts, Cereal, Processed, and Dried Foods: Now I have some questions about nuts, cereals, and processed foods you (your child) might have eaten in the 7 days before your (your child's) illness began. You (your child) could have eaten these either in your home or away from home. As I read each food, please answer as yes, no, may have eaten, or can't remember eating the food.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
◯	◯	◯	◯	1. Pre-packaged peanut butter crackers?
◯	◯	◯	◯	2. Any peanut butter?
				2a. If eaten at home, what was the brand, type/variety: Brand ☐ Creamy ☐ Crunchy ☐ Unknown ☐ Not applicable (did not eat at home)
				2b. If eaten outside the home, where? List name(s) and location(s): ☐ Not applicable (did not eat outside the home)

YES	Maybe	NO	Don't Know	
				3. Peanut butter containing foods (cookies, candies, ice cream, etc.)?
				3a. What was the type, variety, brand: ____ ☐ Unknown
				4. Ground nut butter or spread other than peanut butter (Nutella, almond butter)?

Next I have questions about nuts and seeds you (your child) might have eaten. Remember that these may be used as toppings or mixed into many foods. If you (your child) ate any of the nuts below as part of another food please answer "yes". Did you (your child) eat any of the following:

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
				5. Peanuts?
				6. Almonds (whole, sliced, chopped, etc.)?
				7. Walnuts?
				8. Cashews?
				9. Pistachios?
				10. Hazelnuts or filberts?
				11. Other whole nuts or mixed nuts?
				12. Sunflower seeds?
				13. Sesame seeds?
				14. Tahini, halva, or other product made from sesame seeds?
				15. Hummus?

Section 14 Peanut butter/Nuts/ Seeds Comments. *Please fill in comments/notes from this section in the space provided*

Now I have questions about pre-packaged snack foods and cereals you (your child) might have had in the 7 days before your (your child's) illness began.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
				16. Granola bars, breakfast, power, or protein bars?
				16a. What was the type, variety, brand: ____ ☐ Unknown
				17. Trail mix (or similar product)?
				18. Fruit roll-ups (or similar product)?
				19. Chips or pretzels?
				19a. What was the type, variety, brand: ____ ☐ Unknown
				20. Pre-packaged crackers, cookies, or snack cakes?
				20a. What was the type, variety, brand: ____ ☐ Unknown
				21. Chocolate or chocolate-containing candy?
				21a. What was the type, variety, brand: ____ ☐ Unknown
				22. Cold breakfast cereal?
				22a. What was the type, variety, brand: ____ ☐ Unknown
				23. Hot breakfast cereals like oatmeal, cream of wheat, etc.?
				23a. What was the type, variety, brand: ____ ☐ Unknown

Section 14 Snack foods/Cereal Comments. *Please fill in comments/ notes from this section in the space provided*

And finally I have questions about dried, powdered products and supplements you (your child) might have had in the 7 days before your (your child's) illness began.

YES	Maybe	NO	Don't Know	Did you (your child) eat any:
				24. Dried buttermilk?

○	○	○	○	25. Flavored milk powder (such as chocolate or vanilla)?
○	○	○	○	26. Other powdered milk products?
○	○	○	○	27. Powdered nutritional supplement products?
Section 14 Dried/Powdered foods Comments. *Please fill in comments/notes from this section in the space provided*				

Section 15: We have covered a wide variety of foods, drinks, etc. After answering all these questions are there any other things you (your child) ate or drank in the 7 days before becoming ill that have not been mentioned?

1. Please describe any other foods, drinks, etc. including as much detail as possible regarding type, variety, or brand.

Section 16: <u>**Animal contact and Pets**</u>**:** Now I have some questions about contact with pets or other animals in the 7 days before your (your child's) illness began. This could have been at your home or another home, at a pet store, petting zoo, school, or other location.

YES	Maybe	NO	Don't Know	Did you (your child) visit or go to:
○	○	○	○	1. A petting zoo or farm with livestock like cattle, sheep, goats, etc.?
○	○	○	○	2. Agricultural 'Farm and Feed' stores?
○	○	○	○	3. Pet stores, swap meets, other places where animals/birds were sold or shown?
○	○	○	○	4. County/State fairs, 4-H events, or similar event where animals were present?
○	○	○	○	5. School events, birthday parties, or similar events with animals/pets?
				Did you (your child) have any contact with:
○	○	○	○	6. Dogs or puppies?
○	○	○	○	7. Cats or kittens?
○	○	○	○	8. Baby chicks, duckling, or other baby poultry?
○	○	○	○	9. Live chickens, turkeys, or other adult poultry?
○	○	○	○	10. Turtles or tortoises?
○	○	○	○	11. Snakes?
○	○	○	○	12. <u>Frozen</u> mice, rats, or similar pet food for snakes?
○	○	○	○	13. Other reptiles, such as lizards, geckos, etc.?
○	○	○	○	14. Amphibians, such as frogs, toads, or salamanders?
○	○	○	○	15. Water pets in an aquarium (goldfish, aquatic frogs, snails, etc.)?
○	○	○	○	16. Rats, mice, gerbils, or hamsters?
○	○	○	○	17. "Pocket" or "exotic" pets (ferrets, pygmy hedgehogs, rabbits, sugar gliders, guinea pigs, etc.)?
○	○	○	○	18. Prepackaged pet food (canned or dry)?
				18a. What was the type, variety, brand: ☐ Unknown
○	○	○	○	19. Pet treats or chews (pig ears, puzzles, rawhide, hooves, etc.)?
○	○	○	○	20. Dried animal droppings or pellets (e.g., owl pellets for science projects)?
Section 16 Comments. *Please fill in comments/notes from this section in the space provided*				

That completes the interview, thank you for taking the time to answer all these questions. Your responses may be helpful in preventing others from becoming sick. *(refer to closing script as needed)*

Prior to submitting or printing this form,

please be sure that all necessary sub-questions (including unknown values for sub-questions) and comments are complete.

Submit by Email | Print Form | Reset Form

APPENDIX C – FOOD EXPOSURES SURVEY QUESTIONNAIRES FOR CHAPTER 6

Food Exposures Survey

Field Version 1

"Hello, this is ______________________ calling for the University of South Carolina. This month the University is conducting a confidential study for the Department of Health and Environmental Control about the different types of foods that people in South Carolina eat, and we'd really appreciate your help and cooperation."

Let me make sure I've dialed the correct phone number ... Is this ________________?

"I am going to ask about a number of different foods that you (your child) may or may not have eaten in the past seven days. The first questions are about meat and poultry. This does not include canned items, but the meat and poultry could have been fresh or frozen or eaten as part of a dish. You (your child) could have eaten these either in your home or outside the home. For each of the following items, please tell we whether or not you have (your child has) eaten it in the past seven days, that is since last __________________. First, what about ….

	YES	NO	DK
1. Whole or cut chicken pieces/parts	1	2	3
2. Ground chicken	1	2	3
3. Breaded chicken products, such as chicken tenders	1	2	3
4. Stuffed, frozen chicken products, such as Chicken Kiev	1	2	3
5. Other frozen chicken products	1	2	3
6. Duck, game hen, or squab	1	2	3
7. Whole or cut turkey pieces or parts	1	2	3
8. Ground turkey	1	2	3

"The next questions are about beef products. In the past seven days, did you (your child) eat any:

9. Beef steaks or roasts	1	2	3
10. Pre-made or pre-formed hamburger patties at home	1	2	3
11. Fresh hamburger patties at home	1	2	3
12. Any other ground beef	1	2	3

"Now I have a few questions about pork, lamb, and other meat products. In the past seven days,
did you (your child) eat any:

	YES	NO	DK
13. Ground pork	1	2	3
14. Pulled pork barbecue	1	2	3
15. Bacon	1	2	3
16. Sausage	1	2	3
17. Hot dogs, corn dogs, polish sausage, kielbasa, or similar foods	1	2	3
18. Any other pork product	1	2	3
19. Lamb	1	2	3
20. Pepperoni	1	2	3
21. Any other Italian meats, such as salami, prosciutto	1	2	3
22. Bologna, pastrami, or corned beef	1	2	3
23. Store-bought, dried meat strips or jerky	1	2	3
24. Pre-packaged deli meats	1	2	3
25. Any other deli-sliced meats not pre-packaged	1	2	3

"Now I have some questions about fish and seafood you (your child) may have eaten in the past seven days. You (your child) may have eaten this at home or away from home. This does not include canned items, but these foods could have been eaten as part of a dish, sauce, or dip. For each of these items, please tell me whether or not you have (your child has) eaten it in the past seven days.

	YES	NO	DK
26. Fresh or fresh-frozen fish	1	2	3
27. Smoked or dried fish	1	2	3
28. Shrimp or prawns	1	2	3

	YES	NO	DK
29. Crab, lobster, or crayfish	1	2	3
30. Oysters	1	2	3
31. Clams, mussels, scallops, or other shellfish	1	2	3
32. Sushi with raw fish or seafood	1	2	3
33. Frozen fish products, such as fish sticks, fish nuggets, etc	1	2	3

"Now I have a few questions about eggs, dairy, and cheese products you (your child) might have eaten in the past seven days. You (your child) could have eaten these either in your home or away from home. In the past seven days, did you (your child) eat any:

	YES	NO	DK
34. Eggs or dishes containing eggs	1	2	3
35. Anything made with raw eggs, such as cookie dough, batter, sauces, and the like)	1	2	3
36. Did you (your child) drink any milk (IF NO, SKIP TO Q.38)	1	2	3
37a. Any raw or unpasteurized milk	1	2	3
37b. Any other dairy, such as soy milk or almond milk	1	2	3
38. In the past 7 days, did you (your child) eat any ice cream or ice cream products	1	2	3

39. Frozen yogurt	1	2	3
40. Yogurt drinks	1	2	3
41. Any other yogurt	1	2	3
42. Pre-packaged, shredded cheese	1	2	3
43. Processed, sliced cheese	1	2	3
44. Block-type cheese, such as cheddar, Swiss, Colby, etc	1	2	3
45. String-type cheese	1	2	3

	YES	NO	DK
46. Cottage cheese	1	2	3
47. Feta cheese (this could have been part of a dish or salad)	1	2	3
48. Blue veined cheese, such as gorgonzola or bleu cheese	1	2	3
49. Fresh or dried parmesean, romano, or similar cheese	1	2	3
50. Cheese from raw/unpasteurized milk, for example, homemade, farm-fresh, or sold door-to-door	1	2	3
51. Queso fresco or queso blanco	1	2	3
52. Homemade mexican-style soft cheese	1	2	3

"Now I have some questions about fresh fruits, not canned, cooked, or frozen that you (your child) might have eaten in the past seven days. These foods could have been eaten either in the home or away from home. In the past seven days, did you (your child) eat any:

	YES	NO	DK
53. Apples	1	2	3
54. Fresh lemon or lime - this includes garnishes on a drink	1	2	3
55. Strawberries	1	2	3

56. Raspberries	1	2	3
57. Blueberries	1	2	3
58. Blackberries	1	2	3
59. Cherries	1	2	3
60. Any other fresh berries	1	2	3
61. Cantaloupe	1	2	3
62. Honeydew melon	1	2	3
63. Watermelon	1	2	3

	YES	NO	DK
64. Precut melon or melon salad	1	2	3
65. Any other melon	1	2	3
66. Pineapple	1	2	3
67. Mango	1	2	3
68. Any other tropical fruit, such as kiwi, papaya, guava, pomegranate, and the like	1	2	3
69. Other dried fruit	1	2	3
70. Apple juice - not from concentrate	1	2	3
71. Orange juice - not from concentrate	1	2	3

"The next questions are about fresh vegetables you (your child) might have eaten raw or uncooked in the past seven days. These foods could have been eaten either in the home or away from home. This does not include canned items, but these foods could have been eaten alone or as part of a dish. We are only interested in vegetables that are not grown at home. In the past seven days, did you (your child) eat any:

	YES	NO	DK
72. Red round tomatoes at home	1	2	3
73. Roma tomatoes at home	1	2	3

74. Cherry tomatoes at home	1	2	3
75. Grape tomatoes at home	1	2	3
76. Vine-ripe or sold on the vine tomatoes at home	1	2	3
77. Other fresh tomatoes at home	1	2	3
78. Any tomato (including slices on sandwiches) eaten away from home	1	2	3
79. Fresh salsa or pico de gallo - not from a jar	1	2	3
80. Guacamole	1	2	3
	YES	NO	DK
81. Pre-packaged fresh leafy greens	1	2	3
82. Loose fresh leafy greens	1	2	3
83. Pre-packaged iceberg lettuce	1	2	3
84. Loose iceberg lettuce	1	2	3
85. Pre-packaged romaine lettuce	1	2	3
86. Loose romaine lettuce	1	2	3
87. Pre-packaged fresh spinach	1	2	3
88. Loose fresh spinach	1	2	3
89. Other leafy greens, such as red butter lettuce, radiccio, spring mix, and the like	1	2	3

"Now I have questions about herbs and sprouts you (your child) may have eaten in the past seven days. Remember these could have been part of a dish such as pesto, salsa, sauces, and the like. We are interested in fresh herbs, not dried or bottled herbs. Did you (your child) eat any:

YES NO DK

90. Fresh basil	1	2	3
91. Fresh cilantro	1	2	3
92. Other fresh herbs, such as parsley, sage, thyme, and the like	1	2	3
93. Alfalfa sprouts	1	2	3
94. Other sprouts, such as bean, clover, broccoli sprouts, daikon radish, and the like	1	2	3

"Next I have a few questions about other fresh vegetables you (your child) may have eaten in the past seven days. What about …"

	YES	NO	DK
95. Fresh chile/chili peppers, such as jalapenos or serranos	1	2	3
96. "Mini" carrots – these are often peeled and sold in a sealed bag	1	2	3
7. Raw onions - white, yellow, red or purple	1	2	3
98. Raw green onions or scallions	1	2	3
99. Did you (your child) eat any frozen pot pies	1	2	3

"The next questions are about nuts, cereal, processed, and dried foods. What about .."

	YES	NO	DK
100. Pre-packaged peanut butter crackers	1	2	3
101. Any peanut butter (IF NO, SKIP TO Q.104)	1	2	3
102. Creamy peanut butter eaten at home	1	2	3
103. Crunchy peanut butter eaten at home	1	2	3
104. Foods containing peanut butter, such as cookies, candies, or ice cream and the like	1	2	3
105. Ground nut butter or spread other than peanut butter, such as Nutella or almond butter)	1	2	3

"The next questions are about seeds and nuts you (your child) might have eaten. Remember that these may be used as toppings or mixed into many foods. If you (your child) ate any of these nuts as part of another food, please answer "Yes." What about …"

	YES	NO	DK
106. Peanuts	1	2	3
107. Almonds	1	2	3
108. Walnuts	1	2	3
	YES	NO	DK
109. Cashews	1	2	3
110. Pistachios	1	2	3
111. Hazelnuts	1	2	3
112. Hummus	1	2	3

"And what about pre-packaged foods that you (your child) might have had in the past seven days. What about …"

	YES	NO	DK
113. Granola bars, breakfast, power, or protein bars	1	2	3
114. Chips or pretzels	1	2	3
115. Pre-packaged crackers, cookies, or snack cakes	1	2	3
116. Chocolate or candy containing chocolate	1	2	3

"Now I have some questions about contact with pets or other animals in the past seven days. This could have been at your home, at another home, at a pet store, petting zoo, school or other location. Did you (your child) visit or go to …"

	YES	NO	DK
117. A petting zoo with livestock	1	2	3
118. An agricultural 'farm and feed' store	1	2	3

119. A pet store, swap meets, or other places where animals or birds

are sold or shown	1	2	3
120. County or state fairs, 4-H events, or similar event	1	2	3
121. School event, birthday party, or similar event	1	2	3

"In the past seven days did you (your child) have any <u>contact</u> with:

	YES	NO	DK
122. Dogs or puppies	1	2	3
123. Cats or kittens	1	2	3
124. Baby chicks, ducklings, or other baby poultry	1	2	3
125. Live chickens, turkeys, or other adult poultry	1	2	3
126. Turtles or tortoises	1	2	3
127. Snakes	1	2	3
128. Frozen mice, rats, or similar pet food for snakes	1	2	3
129. Other reptiles, such as lizards, geckos, and the like	1	2	3
130. Amphibians, such as frogs, toads, salamanders, and the like	1	2	3
131. Water pets in an aquarium, such as goldfish, aquatic frogs, snails, and the like	1	2	3
132. Rats, mice, gerbils, or hamsters	1	2	3
133. "Pocket" or "exotic" pets, such as ferrets, pygmy hedgehogs, rabbits, guinea pigs, and the like	1	2	3
134. Pre-packaged pet food - canned or dry	1	2	3
135. Pet treats or chews, such as pig ears, puzzles, rawhide, hooves, and the like	1	2	3

“The final questions are about you so that we can see how different people feel about the types of things we’ve been asking.”

136. "What is your (your child’s) age?"

______ CODE EXACT NUMBER OF YEARS (E.G., 45)

96. NINETY-SIX YEARS OF AGE OR OLDER

97. REFUSED

137. "Do you live in an urban, suburban, or rural area of South Carolina?"

1. URBAN (INSIDE CITY LIMITS)

2. SUBURBAN (JUST OUTSIDE CITY LIMITS)

3. RURAL (AWAY FROM A CITY)

4. DK (PROBE: "How would you describe it?")

138. “Are you (Is your child) of Hispanic or Latino origin?”

1. YES
2. NO
3. DON’T KNOW (DO NOT PROBE)

139. "What is your race?" (PROBE BY READING CHOICES IF NECESSARY)

1. BLACK; AFRICAN-AMERICAN

2. WHITE

3. HISPANIC; PUERTO RICAN; MEXICAN OR SPANISH-AMERICAN

4. NATIVE AMERICAN; AMERICAN INDIAN

5. ASIAN; ORIENTAL

6. OTHER (SPECIFY): ________________________________

140. "How many of the persons who currently live in your household are under 18 years of age, including babies and small children?"

______ RECORD NUMBER

7. SEVEN OR MORE

8. DK

141. "Including yourself, how many people age 18 or older are currently living in your household?"

______ RECORD NUMBER

7. SEVEN OR MORE

8. DK

142. "So that we can be sure we're getting a cross-section of all people, I'd like you to estimate your family's total income for 2011, before taxes were taken out. Include wages, social security, welfare and any other income. Into which of the following categories does it fall? As with all of the interview, this information will be strictly confidential. Was it...

(READ CATEGORIES)

01. Less than $5,000

02. $5,000 - 9,999

03. $10,000 - 14,999

04. $15,000 - 19,999

05. $20,000 - 24,999

06. $25,000 - 29,999

07. $30,000 - 34,999

08. $35,000 - 39,999

09. $40,000 - 44,999

10. $45,000 - 49,999

11. $50,000 - 74,999

12. $75,000 - 99,999

13. $100,000 and over

14. REFUSED

15. DON'T KNOW (PROBE: "Just approximately...")

143. "Not counting business lines, cell phones, extension phones, faxes, or modems -- on how many different land line telephone numbers can your household be reached?"

1. ONE
2. TWO
3. THREE
4. FOUR
5. FIVE
6. SIX
7. SEVEN OR MORE
8. DK

144. "And what is your zip code?" RECORD __________

145. RECORD SEX: 1. MALE

2. FEMALE

That's all the questions I have. Thank you for your cooperation.

Food Exposures Survey

Field Version 2

"Hello, this is ______________________ calling for the University of South Carolina. This month the University is conducting a confidential study for the Department of Health and Environmental Control about the different types of foods that people in South Carolina eat, and we'd really appreciate your help and cooperation."

Let me make sure I've dialed the correct phone number ... Is this ________________?

"I am going to ask about a number of different foods that you (your child) may or may not have eaten in the past seven days. The first questions are about meat and poultry. This does not include canned items, but the meat and poultry could have been fresh or frozen or eaten as part of a dish. You (your child) could have eaten these either in your home or outside the home. For each of the following items, please tell we whether or not you have (your child has) eaten it in the past seven days, that is since last __________________. First, what about ….

	YES	NO	DK
1. Whole or cut chicken pieces/parts	1	2	3
2. Ground chicken	1	2	3
3. Breaded chicken products, such as chicken tenders	1	2	3
4. Whole or cut turkey pieces or parts	1	2	3
5. Ground turkey	1	2	3

"The next questions are about beef products. In the past seven days, did you (your child) eat any:

	YES	NO	DK
6. Beef steaks or roasts	1	2	3
7. Pre-made or pre-formed hamburger patties at home	1	2	3
8. Fresh hamburger patties at home	1	2	3
9. Any other ground beef	1	2	3

"Now I have a few questions about pork, lamb, and other meat products. In the past seven days, did you (your child) eat any:

	YES	NO	DK
10. Ground pork	1	2	3
11. Any other pork product	1	2	3
12. Pre-packaged deli meats	1	2	3
13. Any other deli-sliced meats (not pre-packaged)	1	2	3

"Now I have some questions about fish and seafood you (your child) may have eaten in the past seven days. You (your child) may have eaten this at home or away from home. This does not include canned items, but these foods could have been eaten as part of a dish, sauce, or dip. For each of these items, please tell we whether or not you have (your child has) eaten it in the past seven days.

	YES	NO	DK
14. Sushi with raw fish or seafood	1	2	3
15. Frozen fish products, such as fish sticks, fish nuggets	1	2	3

"Now I have a few questions about eggs, dairy, and cheese products you (your child) might have eaten in the past seven days. You (your child) could have eaten these either in your home or away from home. In the past seven days, did you (your child) eat any:

	YES	NO	DK
16. Eggs or dishes containing eggs	1	2	3
17. Anything made with raw eggs, such as cookie dough, batter, sauces, and the like)	1	2	3
18. Did you (your child) drink any milk (IF NO, SKIP TO Q.20)	1	2	3
19a. Any raw or unpasteurized milk	1	2	3
19b. Any other dairy, such as soy milk or almond milk	1	2	3
20. In the past 7 days, did you (your child) eat any ice cream or ice cream products	1	2	3

	YES	NO	DK
21. Pre-packaged, shredded cheese	1	2	3
22. Processed, sliced cheese	1	2	3
23. Block-type cheese, such as cheddar, Swiss, Colby, and the like	1	2	3
24. String-type cheese	1	2	3
25. Cottage cheese	1	2	3
26. Feta cheese (this could have been part of a dish or salad)	1	2	3
27. Blue veined cheese, such as gorgonzola or bleu cheese	1	2	3
28. Fresh or dried parmesean, romano, or similar cheese	1	2	3
29. Cheese from raw/unpasteurized milk, for example, homemade, farm-fresh, or sold door-to-door	1	2	3
30. Queso fresco or queso blanco	1	2	3

"Now I have some questions about <u>fresh</u> <u>fruits</u>, not canned, cooked, or frozen that you (your child) might have eaten in the past seven days. These foods could have been eaten either in the home or away from home. <u>In</u> <u>the</u> <u>past</u> <u>seven</u> <u>days</u>, did you (your child) eat any:

	YES	NO	DK
31. Apples	1	2	3
32. Grapes	1	2	3
33. Pears	1	2	3
34. Peaches	1	2	3
35. Nectarines	1	2	3
36. Apricots	1	2	3
37. Plums	1	2	3
38. Oranges	1	2	3

	YES	NO	DK
39. Grapefruit	1	2	3
40. Tangerines	1	2	3
41. Fresh lemon or lime - this includes garnishes on a drink	1	2	3
42. Strawberries	1	2	3
43. Raspberries	1	2	3
44. Blueberries	1	2	3
45. Blackberries	1	2	3
46. Cherries	1	2	3
47. Cantaloupe	1	2	3
48. Honeydew melon	1	2	3
49. Watermelon	1	2	3
50. Precut melon or melon salad	1	2	3
51. Pineapple	1	2	3
52. Mango	1	2	3
53. Coconut – whole or shredded	1	2	3
54. Raisins	1	2	3
55. Apple juice - not from concentrate	1	2	3
56. Orange juice - not from concentrate	1	2	3

"The next questions are about fresh vegetables you (your child) might have eaten raw or uncooked in the past seven days. These foods could have been eaten either in the home or away from home. This does not include canned items, but these foods could have been eaten alone or as part of a dish. We are only interested in vegetables that are not grown at home.

In the past seven days, did you (your child) eat any:

	YES	NO	DK
57. Red round tomatoes at home	1	2	3
58. Roma tomatoes at home	1	2	3
59. Cherry tomatoes at home	1	2	3
60. Grape tomatoes at home	1	2	3
61. Vine-ripe or sold on the vine tomatoes at home	1	2	3
62. Other fresh tomatoes at home	1	2	3
63. Any tomato (including slices on sandwiches) eaten away from home	1	2	3
64. Fresh salsa or pico de gallo - not from a jar	1	2	3
65. Guacamole	1	2	3
66. Pre-packaged fresh leafy greens	1	2	3
67. Loose fresh leafy greens	1	2	3
68. Pre-packaged iceberg lettuce	1	2	3
69. Loose iceberg lettuce	1	2	3
70. Pre-packaged romaine lettuce	1	2	3
71. Loose romaine lettuce	1	2	3
72. Pre-packaged fresh spinach	1	2	3
73. Loose fresh spinach	1	2	3
74. Cabbage	1	2	3

"Now I have questions about herbs and sprouts you (your child) may have eaten in the past seven days. Remember these could have been part of a dish such as pesto, salsa, sauces, and the like. We are interested in fresh herbs, not dried or bottled herbs. Did you (your child) eat any:

	YES	NO	DK
75. Fresh basil	1	2	3
76. Fresh cilantro	1	2	3
77. Other fresh herbs, such as parsley, sage, thyme, and the like	1	2	3
78. Alfalfa sprouts	1	2	3
79. Other sprouts, such as bean, clover, broccoli sprouts, daikon radish, and the like	1	2	3

"Next I have a few questions about other fresh vegetables you (your child) may have eaten in the past seven days. What about …"

	YES	NO	DK
80. Cucumbers, zucchini or squash	1	2	3
81. Bell peppers – green, red, orange or yellow	1	2	3
82. Fresh chile/chili peppers, such as jalapenos or serranos	1	2	3
83. Celery	1	2	3
84. "Mini" carrots – these are often peeled and sold in a sealed bag	1	2	3
85. Other fresh carrots	1	2	3
86. Other root vegetables, such as radishes, beets, turnips	1	2	3
87. Fresh raw peas – may be shelled or in the pod	1	2	3
88. Broccoli	1	2	3
89. Cauliflower	1	2	3
90. Raw onions - white, yellow, red or purple	1	2	3
91. Raw green onions or scallions	1	2	3

	YES	NO	DK
92. Fresh or dried mushrooms	1	2	3

"Now I have a few questions about frozen foods you (your child) might have eaten in the past seven days. You (your child) could have eaten these either in your home or outside the home. In the past seven days did you (your child) eat any:

	YES	NO	DK
93. Frozen vegetables – in a bag or box	1	2	3
94. Frozen pot pies	1	2	3
95. Frozen pizza	1	2	3
96. Frozen Mexican-style foods, such as burritos and the like	1	2	3
97. Frozen snack foods, like mozzarella sticks, jalapeno poppers, potato skins, or hot pockets	1	2	3
98. Frozen breakfast items, such as waffles, breakfast sandwiches and the like	1	2	3
99. Frozen vegetarian foods such as a garden burger	1	2	3
100. Frozen pre-mixed meals in a bag or box, such as stir fry, pasta meals, and the like	1	2	3
101. Frozen dinners or box entrees	1	2	3

"The next questions are about nuts, cereal, processed, and dried foods. What about .."

102. Pre-packaged peanut butter crackers	1	2	3
103. Creamy peanut butter eaten at home	1	2	3
104. Crunchy peanut butter eaten at home	1	2	3
105. Foods containing peanut butter, such as cookies, candies, or ice cream and the like	1	2	3

	YES	NO	DK
106. Ground nut butter or spread other than peanut butter, such as Nutella or almond butter)	1	2	3

"The next questions are about seeds and nuts you (your child) might have eaten. Remember that these may be used as toppings or mixed into many foods. If you (your child) ate any of these nuts as part of another food, please any "Yes." What about …"

	YES	NO	DK
107. Peanuts	1	2	3
108. Almonds	1	2	3
109. Walnuts	1	2	3
110. Cashews	1	2	3
111. Pistachios	1	2	3
112. Hazelnuts	1	2	3
113. Sunflower seeds	1	2	3
114. Sesame seeds	1	2	3
115. Tahini, halva, or other products made from sesame seeds	1	2	3
116. Hummus	1	2	3

"And what about pre-packaged foods that you (your child) might have had in the past seven days. What about …"

	YES	NO	DK
117. Trail mix or a similar product	1	2	3
118. Fruit roll-ups or a similar product	1	2	3
119. Pre-packaged crackers, cookies, or snack cakes	1	2	3
120. Cold breakfast cereal	1	2	3
121. Hot breakfast cereals like oatmeal, cream of wheat	1	2	3

"And a few questions about dried, powdered products and supplements you (your child) might have had in the past seven days. What about …"

	YES	NO	DK
122. Dried buttermilk	1	2	3
123. Flavored milk powder, such as chocolate or vanilla	1	2	3
124. Other powdered milk products	1	2	3
125. Powdered nutritional supplements	1	2	3

"Now I have some questions about contact with pets or other animals in the past seven days. This could have been at your home, at another home, at a pet store, petting zoo, school or other location. In the past seven days did you (your child) have any contact with …"

	YES	NO	DK
126. Baby chicks, ducklings, or other baby poultry	1	2	3
127. Live chickens, turkeys, or other adult poultry	1	2	3
128. Turtles or tortoises	1	2	3
129. Snakes	1	2	3
130. Frozen mice, rats, or similar pet food for snakes	1	2	3
131. Other reptiles, such as lizards, geckos, and the like	1	2	3
132. Amphibians, such as frogs, toads, salamanders, and the like	1	2	3
133. Water pets in an aquarium, such as goldfish, aquatic frogs, snails, and the like	1	2	3
134. Rats, mice, gerbils, or hamsters	1	2	3
135. "Pocket" or "exotic" pets, such as ferrets, pygmy hedgehogs, rabbits, guinea pigs, and the like	1	2	3
136. Pre-packaged pet food - canned or dry	1	2	3
137. Pet treats or chews, such as pig ears, puzzles, rawhide, hooves, and the like	1	2	3

"The final questions are about you so that we can see how different people feel about the types of things we've been asking."

138. "What is your (your child's) age?"

______ CODE EXACT NUMBER OF YEARS (E.G., 45)

96. NINETY-SIX YEARS OF AGE OR OLDER

97. REFUSED

139. "Do you live in an urban, suburban, or rural area of South Carolina?"

1. URBAN (INSIDE CITY LIMITS)

2. SUBURBAN (JUST OUTSIDE CITY LIMITS)

3. RURAL (AWAY FROM A CITY)

4. DK (PROBE: "How would you describe it?")

140. "Are you (Is your child) of Hispanic or Latino origin?"

1. YES
2. NO
3. DON'T KNOW (DO NOT PROBE)

141. "What is your race?" (PROBE BY READING CHOICES IF NECESSARY)

1. BLACK; AFRICAN-AMERICAN
2. WHITE
3. HISPANIC; PUERTO RICAN; MEXICAN OR SPANISH-AMERICAN
4. NATIVE AMERICAN; AMERICAN INDIAN
5. ASIAN; ORIENTAL
6. OTHER (SPECIFY): ________________________________

142. "How many of the persons who currently live in your household are under 18 years of age, including babies and small children?"

______ RECORD NUMBER

7. SEVEN OR MORE

8. DK

143. "Including yourself, how many people age 18 or older are currently living in your household?"

______ RECORD NUMBER

7. SEVEN OR MORE

8. DK

144. "So that we can be sure we're getting a cross-section of all people, I'd like you to estimate your family's total income for 2011, before taxes were taken out. Include wages, social security, welfare and any other income. Into which of the following categories does it fall? As with all of the interview, this information will be strictly confidential. Was it...

(READ CATEGORIES)

01. Less than $5,000

02. $5,000 - 9,999

03. $10,000 - 14,999

04. $15,000 - 19,999

05. $20,000 - 24,999

06. $25,000 - 29,999

07. $30,000 - 34,999

08. $35,000 - 39,999

09. $40,000 - 44,999

10. $45,000 - 49,999

11. $50,000 - 74,999

12. $75,000 - 99,999

13. $100,000 and over

14. REFUSED

15. DON'T KNOW (PROBE: "Just approximately...")

145. "Not counting business lines, cell phones, extension phones, faxes, or modems -- on how many different land line telephone numbers can your household be reached?"

1. ONE
2. TWO
3. THREE
4. FOUR
5. FIVE
6. SIX
7. SEVEN OR MORE
8. DK

146. "And what is your zip code?"   RECORD __________

147. RECORD SEX: 1. MALE
                   2. FEMALE

That's all the questions I have. Thank you for your cooperation.

APPENDIX D – A GUIDE TO RANDOM FORESTS

D.1 WHAT ARE RANDOM FORESTS?

Random Forests are one of the most powerful, fully automated, machine learning techniques. With almost no data preparation or modeling expertise, analysts can effortlessly obtain surprisingly effective models. Random Forests is a tool that leverages the power of many decision trees, judicious randomization, and ensemble learning to produce accurate predictive models, insightful variable importance rankings, missing value imputations, novel segmentations, and laser-sharp reporting on a record-by-record basis for deep data understanding.[1]

D.2 HOW DO ENSEMBLE MODELS WORK?

Take a large collection of individually imperfect models. The one-off mistakes made by an individual model will most likely not be made by the rest of the models. If we average the results of all these models, we can sometimes find a superior model from their combination than any of the individual parts. That is how ensemble models work, they grow many different models, and let their outcomes be averaged or voted across the group.[2] Let's build a very small ensemble of three simple decision trees to illustrate:

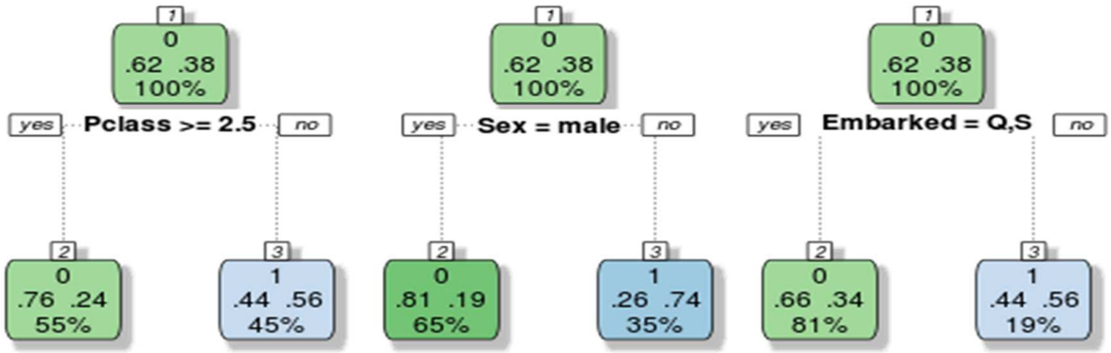

Figure D.1 Small Decision Tree - This tree shows probabilities of being in each split. The first tree describes whether or not the passenger is in first class. Tree two shows whether or not the passenger's gender is male and tree three shows whether or not the person embarked on their flight.

Each of these trees make their classification decisions based on different variables. So let us imagine that a female from Columbia took a first class flight. Trees one and three would vote for this scenario, but tree two votes that she is male. There is a vote of 2 to 1 in this situation, so we would vote for success that she is a female passenger from Columbia who took a first class flight. Random forests grow much deeper trees that those above. In fact, the default behavior is to grow each tree as far as possible. However, since the formulas for building a single decision tree are the same every time, some source of randomness is required to make these trees different from one another.

D.3 ADDING RANDOMNESS TO THE MODELS

In Random Forests a new random subset of predictors in each split of a tree is selected. A completely different set of variables may be chosen at each split. If the tree grows large, then by the end of the process many variables have had a chance to influence the tree.[1] If we always search all predictors in every node of every tree, we are building bagger models that are typically not so impressive in their performance. The performance will usually improve if we search fewer than all the variables in each node, meaning restricting to a random subset. This is called bagging. Bagging takes a randomized sample of the rows in your subset, with replacement.[2] For example, let us simulate the process using the sample function in R on 10 rows with replacement.[3]

```
> sample(1:10, replace = TRUE)
 [1] 1 3 7 2 6 2 3 4 1 9
```

Figure D.2 Sampling 10 Rows in R with Replacement - This figure shows the sample function run in R with replacement

As you can see, we would still have 10 rows to work with, but rows 1, 3, and 2 are each repeated twice, while rows 5, 8, and 10 are excluded completely. Every time you run the simulation, you will receive a different set of samples. On average, around 37% of the rows will be left out of the bootstrapped sample.[2] With these repeated and omitted rows, each decision tree grown with bagging would evolve slightly differently.

Random forests actually gets past this limitation by taking only a random subset of variables from the total, typically the square root of what is available.[1] So, using our previous example of 10 variables, random forests would choose a subset of 3 for each and every split of the decision tree. This way, each variable has the same opportunity to be in the model.

Through the addition of randomness, the ensemble contains a collection of unique trees that all make their classifications differently. Since each tree is grown out fully, they each overfit, but in different ways. Thus, the aforementioned one-off mistakes one model makes will be averaged out over all models to give the best possible modeling.

D.4 VARIABLE IMPORTANCE

There are two types of variable importance included in random forests. It is a method to measure the relative importance of any predictor based on measuring the damage that would be done to our predictive models if we lost access to true values of a given variable.[1] Accuracy tests to see how worse the model performs without each variable. GINI essentially measures how pure the splits are at the end of the tree.[2]

To simulate losing access to a predictor, values are randomly scrambled in the data. That is, the value belonging to a specific row of data is moved to another row.[1] This is done one predictor at a time and the loss in accuracy is measured. Random forests scrambles the data for each predictor being tested in every tree in the forest, which removes the dependence on luck of the draw predictions. For example, if a predictor is scrambled 500 times in front of 500 trees, the results should be highly reliable.[1] While that measures accuracy, there is another measure, GINI. GINI is based on the actual role of a predictor and offers an alternative importance assessment based on the role the predictor plays in the data. It is a measure of how often a chosen predictor would be incorrectly classified if it was classified at random based on the subset of data chosen at each tree.[4] GINI then calculates each predictor importance as the sum over the number of splits across all trees, giving a fast variable importance that is often very consistent with the permutation importance measure.

D.5 STRENGTHS/WEAKNESSES

There are plenty of advantages to using random forests. First, it is one of the most accurate machine learning algorithms available. It can also run efficiently on large datasets and handle thousands of input variables without variable deletion.[5] Random forests also gives estimates of what variables are important in classification. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.[5] Random forests is also easily parallelized. It is an ensemble of independently built decision trees. This means that no tree in the ensemble depends on any other tree. This allows for incredibly fast analysis.[1] It also offers an experimental method for detecting variable interactions. Also, generated forests can be saves for future use on other data.[5]

However, there are overfitting problems with decision trees. Also, for data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data. But, if we grow a lot of trees and have them vote on the outcome, we can get beyond this limitation.

D.6 CONCLUSION

As stated previously, random forests are one of the most powerful, fully automated, machine learning techniques. Random Forests are incredibly fast to build and even faster in its predictions, especially since it is easily parallelizable.1 It is well suited for the analysis of complex data structures in datasets with possibly millions of variables. Random forests is a crucial element for any data scientist.

D.7 WORKS CITED

1. Salford Systems. Random Forests for Beginners. http://info.salford-systems.com/an-introduction-to-random-forests-for-beginners. Published 2014. Accessed 19 Feb 2017.
2. Stephens T. Titanic: Getting Started With R - Part 5: Random Forests. http://trevorstephens.com/kaggle-titanic-tutorial/r-part-5-random-forests/. Published 2014. Accessed 28 Feb 2017.
3. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
4. Strickland J. "9. Decision Tree Learning." Data Analytics Using Open-Source Tools. N.p.: Lulu.com, 2016. 201.
5. UC Berkeley. Random Forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm. Published 2001. Accessed 20 Sep 2014.