A Comparative Analysis Between Three Spatio-Temporal Scan Statistics for Outbreak Detection and Antimicrobial Resistance

A Comparative Analysis between Three Spatio-Temporal Scan Statistics for Outbreak Detection and Antimicrobial Resistance

Marissa Bokhari

A Thesis in the Field of Biology

For the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2020

Abstract

Antimicrobial resistance is a growing problem throughout the world. Government agencies and other global health organizations have created software for tracking emerging microbial threats on a facility, state-wide and international scope. Epidemiological software uses statistical algorithms to discern between aberrations of resistance in a given location and time, and whether these "clusters" can be attributed to a specific cause versus chance alone. The retrospective spatiotemporal scan statistic software, SaTScan, has diverse applications. In this project we demonstrate the use of SaTScan version 9.8 within the free epidemiological software WHONET. Microbiology data collected from 2000-2006 at Brigham and Women's Hospital was used for retrospective analysis. Three algorithms: the space-time continuous uniform, the space-time discrete Poisson and the space-time permutation were run on the same data set with limited parameter adjustments per model. Three variables were considered within each model: maximum spatial size within the population at risk, baseline days, maximum temporal size and number of Monte Carlo simulations. As demonstrated by this study, we found that there was a significant difference in cluster detection between the three probability models. Furthermore, we demonstrate the effects of varying values when defining the spatiotemporal scanning window.

The findings of this study allow epidemiologists and computational biologists alike in better understanding the parameters of these algorithms, and create a more precise narrative of outbreaks in antimicrobial resistant clusters. Further prospective validation of these results would give a better understanding of the accuracy and clinical importance for the clusters found. This could be utilized world-wide for other WHONET users, and can guide public health policies in the event of an outbreak or epidemic.

Acknowledgments

I would like to thank my thesis director and mentor, Dr. John Stelling, for teaching me about WHONET and beyond. I would also like to thank Adam Clark, Dr. Martin Kulldorff, Dr. Thomas F. O'Brien and Dr. Meghan Baker and her team for their time and guidance. Lastly, I want to thank my family, friends, and my dog Zelda for their unwavering support.

At the time of this project, the novel Coronavirus (CoVid-19) pandemic has swept the globe, resulting in economic collapse, social isolation, and thousands of lives lost. I want to acknowledge my mentors and teachers working on the frontline in the Department of Medicine, Division of Infectious Disease at Brigham and Women's Hospital. I hope this work provides some clarity in these uncertain times.

Table of Contents

Abstract……iii

Acknowledgments…….....iv

List of Tables……..……..vii

List of Figures……..…….....viii

Chapter I Introduction……...……1

Definition of Terms…….....4

Background of the Problem……...…….....8

19th Century Epidemiology……...8

Modern Epidemiology…….....13

Introduction to Scan Statistics……...18

Question and Hypothesis……...27

Implication of Research……...……..28

Chapter II Materials and Methods……...……..……..30

Data for Analysis...…….....31

Analytical Tools……...……...……31

Research Limitations…….....33

A Note on Monte Carlo Simulations and Gumbel Locally Adjusted Spatial Scan Statistics……33

Chapter III Results……..…….....38

Baseline Comparisons……....38

Parameter Hypothesis Testing……..…..40

Chapter IV Discussion……..……46

Future Work……..47

Conclusion……...……49

References……50

Appendix……..……56

List of Tables

Table 1 Summary of Isolate Frequencies……....57

Table 2 Data Analyses Codes (MSPR, BLD, MTS)……..58

Table 3 Experimental Design Rationale…….59

Table 4 Clusters Detected by Models with Baseline Values (Macros: CU1, PD1 and STP1)……..60

Table 5 Parameter Hypothesis Testing: Maximum Spatial Size in Population at Risk (MSPR) (Macros: CU2, PD2 and STP2) …….63

Tabe 6 Parameter Hypothesis Testing: Baseline Days (BLD)(Macros: CU3, PD3 and STP3)……..65

Table 7 Parameter Hypothesis Testing: Maximum Temporal Size (MTS) (Macros: CU4-16, PD4-16 and STP4-16)……67

List of Figures

Figure 1 John Snow (1854) Cholera Map, SoHo London……11

Figure 2 Shiode et. al (2015) Map of Victim Locations with Additional Records……..12

Figure 3 Shiode et. al (2015) Spatiotemporal Map of Cholera Deaths……13

Figure 4 Gwitria et. al (2018) Distribution of HIV/AIDS and Malaria Clusters…….15

Figure 5 Kulldorff, Hwang and Konty (2009) Geographical Distribution of Birth Weight in New York City zip codes in 2004……..22

Figure 6 Weaver et. al (2015) Comparison of Observed Versus Expected Hospitalization Rate Frequencies……24

Figure 7 Costa and Kulldorff (2014) Cylindrical Cluster Candidate ……..27

Figure 8 Baseline Space-Time Permutation Metaward Only Macro……...56

Figure 9 Baseline Space-Time Permutation Metaward by Resistance Macro…….56

Chapter I

Introduction

Antimicrobial resistance (AMR) remains a dire threat in modern-day healthcare, and requires vigilant surveillance. Resistance occurs when a microbe obtains a new gene or loses sensitivity to a specific antibiotic. The resistance mechanism of action can vary widely from interference of wall synthesis, inhibition of protein or nucleic acid synthesis, interruption of a metabolic pathway, or even disruption of bacterial membrane structure as seen in polymyxins and daptomycin (Tenover, 2006). As the use of antimicrobial agents increases, bacterial pathogens have responded with more creative and complex methods of resistance. A similar dynamic has been coined "the Red Queen Hypothesis", in which a parasite-host relationship coevolves through perpetual negative frequency-dependent selection, resulting in infinite alternating cycles of dominance and population abundance (Rabajante et al, 2015). In relation to antimicrobial resistance, Anzia and Rabajante (2018) demonstrated through simulations that uninterrupted application of an antibiotic of high-level effectiveness (>85%) was needed in order to escape the endless Red Queen dilemma and suppress the evolution of resistant pathogen phenotypes. However, it was also noted that interruption or discontinuation of the antibiotic could result in the return of potent parasites and sudden extinction of hosts.

Many global and government agencies have practiced outbreak recognition and antimicrobial surveillance with algorithms and software that addresses various dimensions of this growing problem. In the early 2000's, the Centers for Disease Control and Prevention (CDC)

introduced The Early Aberration Reporting System (EARS) cluster detection algorithms as a strategy for prompt recognition of emerging threats, including AMR, bioterrorism and other modern public health crises. EARS was also applied to syndromic surveillance for large crowd events: following the terrorist attacks of September 11, 2001, the algorithm was modified for use as a standardized monitoring system beyond infectious disease outbreaks (Hutwagner, Thompson, Seeman & Treadwell, 2003). The algorithm itself focuses on aberration detection, which is defined by a change in expected frequency or distribution of a health-related event when compared to historical or recent data. The criterion for distribution comparison is assigned by the researcher depending on the study of interest. In this project, we delve into the nuances of distribution parameters using algorithms similar to EARS that are better integrated with infectious disease epidemiology.

The free spatiotemporal scan statistic software SaTScan was originally created in 1997 by Dr. Martin Kulldorff and Information Management Services Inc. and has since proliferated into a helpful tool for assessing scan statistics across various platforms, including R and SAS. SaTScan-related software is also integrated within the free database software WHONET developed in 1989 by Dr. John Stelling and the World Health Organization Collaborating Centre for Surveillance of Antimicrobial Resistance, based at the Brigham and Women's Hospital in Boston, Massachusetts. WHONET software is installed in over 120 countries worldwide for national microbiology laboratory reporting. Along with its complementary data capture and conversion software, BacLink, it is able to extract information from a wide variety of hospitals, laboratories and other facilities. Within the context of WHONET, SaTScan assists in early detection of resistant microbial patterns.

Scanning statistics aid greatly in determining potential clusters of antimicrobial resistance or outbreaks. However, the parameters and output variation between each test are not succinctly defined. Numerous studies investigate scanning shapes and windows, and even the optimization of maximum recorded cluster size (Kim & Jung 2017), yet there remains a dearth of understanding in the differences and potential downfalls when choosing one algorithm over another. A clue towards this deficit was demonstrated by Ward and Farnsworth (2009) with data of highly pathogenic avian influenza subtype H5N1 from a village in Romania over the course of nine months. When comparing the space-time permutation to the space-time discrete Poisson model, the number of clusters detected by both was approximately the same, but the permutation model failed to recognize clusters present in all areas studied.

This paper focuses on three particular spatiotemporal algorithmic models used by SaTScan software to understand the crucial differences between each: the space-time continuous uniform model, the space-time discrete Poisson distribution model and the space-time permutation model. The three models investigated in this project involve a cylindrical scanning window with an elliptical or circular base, with the option to define a non-Euclidian distance metric. The distance metrics for our data are dictated by hospital ward locations, which we will elaborate on in chapter three. At first glance, one might argue that running all three models would provide sufficient insight on a data set. A critical mistake in analyses occurs when running more than one algorithm without adjusting for all window sizes. As Han et. al (2016) noted, it is dishonest to run a statistic multiple times with varying maximum spatial window sizes and then simply selecting the cluster with the lowest p-value. This error results in biased p-values and distracts from perhaps more relevant but smaller subclusters. The authors of SaTScan suggest

using an advanced feature to rerun the analyses while holding the same fixed maximum spatial window size while varying the maximum reported cluster sizes. Generally, cluster definition should have a default maximum spatial window size less than 50% of the population risk, with outbreak detection of rare organisms scanning up to 90% of a database. The default assessment understandably holds the maximum spatial window size (MSWS) around 50% or less, with maximum reported cluster size (MRCS) values at increments such as 5, 10, 20 and 50% (Kim & Jung, 2017). However, there has been no investigation in adjusting these values within three different algorithms. Through different statistical methods and a singular data set, this paper exemplifies the risks of each model through exploratory analysis and offers suggestions for future analyses.

Definition of Terms

> "Acquired resistance": refers to bacterium that were previously susceptible but gained resistance to a singular antibiotic through mutation or acquisition of DNA.
>
> "Carrier": An individual who harbors an infectious pathogen without displaying symptoms.
>
> "Cluster": an aggregate of similar things that occur in close spatiotemporal proximity. In scan statistics, a cluster indicates an aggregate of disease or other health-related cases with an unusually high incidence rate, however this may be an epidemiologically irrelevant spurious correlation.
>
> "Community-acquired infections": in contrast to HAIs, infections from these pathogens are obtained from a patient's community and environment outside of the healthcare setting. Some examples include seasonal influenza, *Streptococcus pneumoniae*, *Legionella pneumophila*, Norovirus and HIV/HCV infections.
>
> "Continuity correction": an adjustment used in probability theory for improved accuracy. Made when a discrete distribution is approximated by a continuous distribution.
>
> "Continuous uniform algorithm": in the context of this project, refers to the Normal model used by SaTScan v. 9.6 for geospatial scanning for different types of continuous data, such as lead levels in children or low birth weight. This model runs on the prior

assumption that data has a Gaussian distribution, and as such the Null hypothesis states that any aberrations detected come from the same distribution (SaTScan, 2020).

“CoVid 19”: A novel infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Highly contagious, affects both humans and animals. Symptoms are similar to pneumonia. Declared a pandemic in March 2020 by the World Health Organization (WHO).

“Ebola Virus”: is a rare infectious disease that results in severe bleeding and organ failure known as Ebola haemorrhagic fever (EHF). Has a 25-90% mortality rate, and is treated with palliative care.

“Epidemic”: when an unexpected increase in disease cases occurs in a short period of time. More severe than “sporadic” or “endemic”, but less severe than “pandemic”.

“Gaussian distribution”: also known as a “normal distribution” or “bell curve”, it is a type of continuous probability distribution for any given real-valued variable. In a Gaussian distribution, the μ equals the expected value of distribution, with σ as its standard deviation, as generally defined by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

“GIS (Geographic Information System)”: a framework for capturing, managing, analyzing and presenting spatial information and geographic data. GIS structures use spatio-temporal location as the key index variable for all other related information within a given dataset.

“Gumbel approximation”: an adjustment made to a scan statistic to reduce the “local multiplicity problem” where clusters overlap due to neighborhoods or spatiotemporal classifications. Gumbel approximations assess the distribution of each cluster in relation to its geographic locale.

“Horizontal gene transfer”: also known as “lateral gene transfer” occurs when genetic material is passed “vertically” between bacteria without the transmission of DNA from parent to offspring, as seen in reproduction or binary fission. This method of gene transfer can contribute to rapid exchange of resistance genes between pathogens.

“Hospital-acquired infections (HAIs)”: Also known as “nosocomial infections” are caused by viral, bacterial or fungal pathogens that infect patients when receiving medical treatment within a healthcare facility, but are absent at the time of admission. Some examples include bloodstream infections caused by central lines, catheter-associated urinary tract infections, surgical-site infections, ventilator-associated pneumonia and *Clostridioides difficile* infections (Monegro & Regunath, 2020).

"Hospital Infection Control program": series of steps and precautionary measures to prevent and contain contagious hospital-acquired infections. Examples include hand-washing and employee hygiene, proper waste disposal, personal protective equipment (PPE) and cleaning/disinfecting practices.

"Incidence rate": in epidemiological terms, incidence rate is a measure of the probability of the occurrence of a medical condition happening within a population during a particular interval of time. It can determine the likelihood of being diagnosed with a disease during a given period of time.

"LLR (Log-likelihood ratio)": a test assessing likelihood ratios between two statistical models to determine goodness of fit. The ratios are found by mathematical optimization of parameter space and by imposing any given parameter constraint.

"Monte Carlo hypothesis testing": is a computer-heavy method of testing the robustness of a statistical analysis. It involves setting a specific benchmark result for a specific model, and then running a repetitive analysis chain of student *t*-test distributions for v degrees of freedom for a predetermined hypothesis (Elias et. al 2006).

"Multidrug-resistant phenotype": differs depending on species. As defined by the European Centre for Disease Prevention and Control (ECDC), a multidrug-resistant phenotype (MDR) is defined as a microbe with acquired non-susceptibility to at least one agent in three of more antimicrobial categories. An extensively drug-resistant phenotype (XDR) is a bacterial isolate with susceptibility to only one or two antimicrobial categories. Pandrug-resistant phenotypes (PDRs) are defined as microbes resistant to all agents of all antimicrobial categories. Due to the limited number of novel antimicrobial agents in development, the recent growth in antimicrobial resistance has become a major global threat (Magiorakos et. al, 2012).

"Multiple integral": the definite integral of a function for more than one variable (i.e. $f(x,\ y)$ or $f(x,\ y,\ z)$. In calculus, an integral is a formula that assigns numbers descriptions to metrics such as area, volume or displacement.

"MRCS (Maximum Reported Cluster Sizes)": in SaTScan is programmed at a default of 50%, meaning each cluster detected is more significant than at least half of the rest of the data in consideration. For rare organisms/phenotypes, or for wards with low data volumes, an MRCS of 90-100% would be more appropriate.

"MRSA (Methicillin-resistant *Staphylococcus aureus*)": MRSA are *S. aureus* that have acquired and incorporated a staphylococcal cassette chromosome encoding methicillin resistance (SCC*mec*) into the genome, conferring resistance to β-lactam antibiotics. MRSA is seen in both hospital- and community-acquired infections (Lindsay 2013).

“Pandemic”: a disease affecting many people and multiple countries. Most often used to describe a contagion that is prevalent worldwide for which all people have risk of exposure.

“Patient zero”: a term referring to the first infected person in a major epidemic. The illness is usually in its least mutated form.

“Discrete Poisson distribution algorithm”: in the context of this project, refers to the Poisson-based model used by SaTScan v. 9.6, which assumes the number of events in a given geographical area are Poisson-distributed according to a known underlying population at risk (SaTScan, 2020).

“Population-at-Risk”: is defined as the population of people exposed to the occurrence of a vital event, and can be quantified by dividing the population attributable risk (PAR) by incidence in the total population and then multiplying by 100 for a percentage.

“Recurrence interval”: the inverse of the *p*-value. In epidemiological terms, it is the probability that any given event will be equaled or exceeded during any given time (i.e. the probability of diagnosing a higher number of strep throat cases within the next year compared to the past five.)

“Selective reporting”: refers to the laboratory practice of suppressing antibacterial susceptibility results. This method of antimicrobial stewardship provides only 5-6 antibiotics to prescribers to offset the rise of excessive antibiotic use and acquired resistance of virulent pathogens.

“Sensitivity”: in epidemiological terms, this refers to the proportion of actual positives that are correctly identified. For example, in this project, a model with high sensitivity is one with a high percentage of clusters that are correctly identified as an outbreak.

“Space-time permutation algorithm”: in the context of this project, refers to the space-time permutation model used by SaTScan v. 9.6, which exclusively uses case data to determine geographic clustering. This model does not require population-at-risk data and can be used for early detection of a disease outbreak with only the number of cases available. Two variants of this algorithm allow both prospective and retrospective geospatial scanning for aberration detection of any given location and size by examining potential one-day and multi-day outbreaks (SaTScan, 2020).

“Spatial Scan Statistics”: Statistical method to relate the clustering of randomly positioned points. It is used to detect spatial or space-time clusters, to test whether a disease is randomly distributed over space, over time, or over space and time, and to evaluate the statistical significance of disease clusters (SaTScan, 2020).

“Specificity”: in epidemiological terms, this refers to the proportion of actual negatives that are correctly identified. For example, in this project, a model with high specificity is

one with a high percentage of clusters that are correctly identified as insignificant for hospital infection control.

"VRE (Vancomycin-resistant enterococci)": A species of enterococci that acquires vancomycin resistance by obtaining the mobile *vanA* or *vanB* gene complex. The most common strains are *Enterococcus faecalis* and *Enterococcus faecium*. VRE is among the top causes of nosocomial infections worldwide (Crank & O'Driscoll, 2015).

Background of the Problem

19th Century Epidemiology

Quantification and containment of outbreaks have played a key role in the history of sanitation and preventive medicine. One of these earlier efforts occurred during the revitalization of Paris in the late 19th century. Baron Georges-Eugene Haussmann was a government official and urban planner commissioned by Emperor Napoleon III to renovate the city's layout. At the time, Central Paris was riddled with poverty and disease due to the confinements of a medieval layout. Instead of expanding the city limits, property developers simply built new floors upon already existing structures. The creation of slums, where cramped confines and resistance towards converging waste removal into the sewer system, resulted in rampant cholera and typhoid epidemics (Gandy, 1999). The city's lead engineer A. A. Mille had successfully used sewer water for fertilization of a local field of crops, but the idea of feces mixed into irrigation was unsurprisingly met with resistance. It was not until Hausmann' initiative (1894) that integration of human waste and sewage was mandated by law, despite landlords' reluctance to move away from the cesspool system.

Sometimes the pursuit of patient zero had greater social implications, as seen in the cautionary tale of Mary Mallon. Also known as "Typhoid Mary", Mallon ignited an outbreak starting in March 1907 in which nearly 3,000 New Yorker City residents became infected with

Salmonella Typhi (Marineli et. al, 2013). Mary herself was an Irish immigrant who worked as a cook, and was an asymptomatic carrier for the disease. The public health investigators were unable to determine the mechanism of disease transmission as Mary continued to work, thus contaminating all the food she prepared. While initially believed to stem from soft clams, George Sober, a sanitary engineer, was hired to stalk and obtain fecal, urine or blood samples from Mary after noticing a pattern in the families she had been serving. She was eventually forced to provide stool samples, for which 120/163 tested positive for the bacteria. Mary was then quarantined into a cottage for two-years. She was never provided with an explanation of the disease of her role as a "carrier", and was unsuccessfully treated with a range of antibiotics, laxatives and Brewer's yeast. She was even offered a gallbladder removal. However, since her medical condition and its consequences were not adequately explained to her, upon returning to society she resumed her role as a chef. Within three months Mary had infected 25 people at Sloane Maternity in Manhattan, including doctors, nurses and patients. After two of the infected had died and she had been sent back to North Brother Island, the term "Typhoid Mary" became a household name and even appeared in medical textbooks. The negative stigma towards her was an unfortunate consequence of premature bioethical standards and an unsettling example of social attitudes towards disease carriers. The case of Mary Mallon also demonstrates the necessity for data-driven surveillance of disease outbreaks.

Before sophisticated algorithms and geospatial tools, public health officials calculated and plotted comparative mortality ratios on maps to monitor the spread of disease. The focus on cartographic analysis began most notably in Soho, London in 1854, when British physician John Snow successfully traced and terminated the source of a large cholera outbreak. Dr. Snow plotted

578 cholera victims' households on a base map and traced an equidistance line around areas served by different water pumps within the neighborhood. He was able to pinpoint the Broad Street pump as the root of the epidemic from the high concentration of cases from the area it services, as seen in Figure 1. His discovery halted the outbreak and confirmed his suspicion that the disease stemmed from a water source, while his contemporaries believed it to be airborne. While Snow was a pioneer in sanity science, his findings were almost entirely based on cross-sectional data. Shiode et. al (2015) replicated Snow's experiment with historical records (Figure 2) using modern scan statistics to include population-at-risk and progression of the epidemic over time (Figure 3). The study used NetScan, a software for determining statistically significant concentrations along street along street networks. NetScan applied the Kernel Density Estimation (KDE) method to estimate a density surface across the area studied using known recorded values. High mortality rates observed around the Broad Street pump and the absence of a clear spatiotemporal pattern were consistent with the waterborne nature of cholera. Furthermore, application of KDE showed a distinct tendency of distance-decay in mortality rate with increasing distance from the Broad Street pump.

Figure 1. John Snow's cholera map, SoHo London, 1854

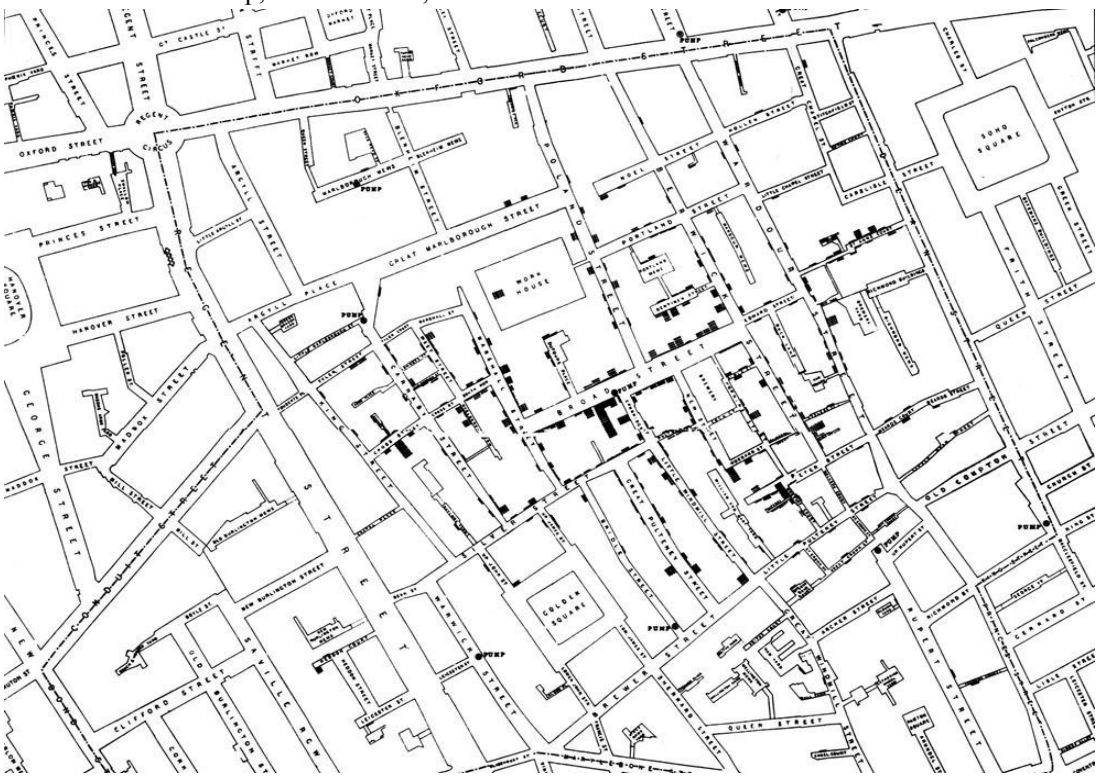


Figure 2. Shiode et. al (2015) map of victim locations with additional records

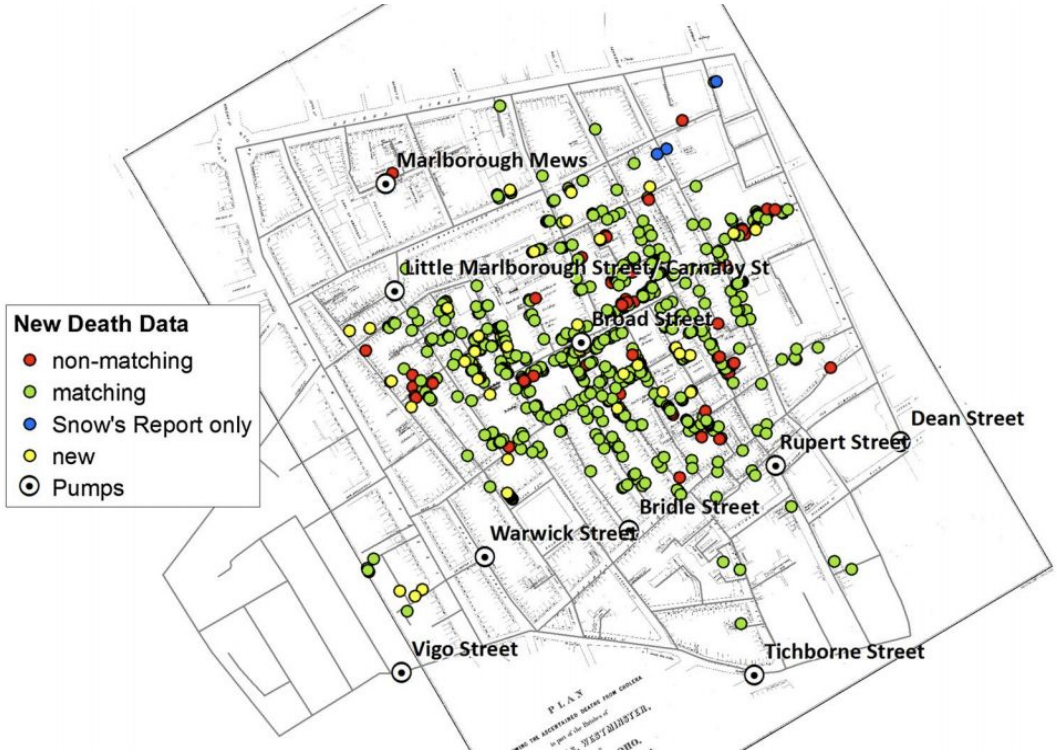
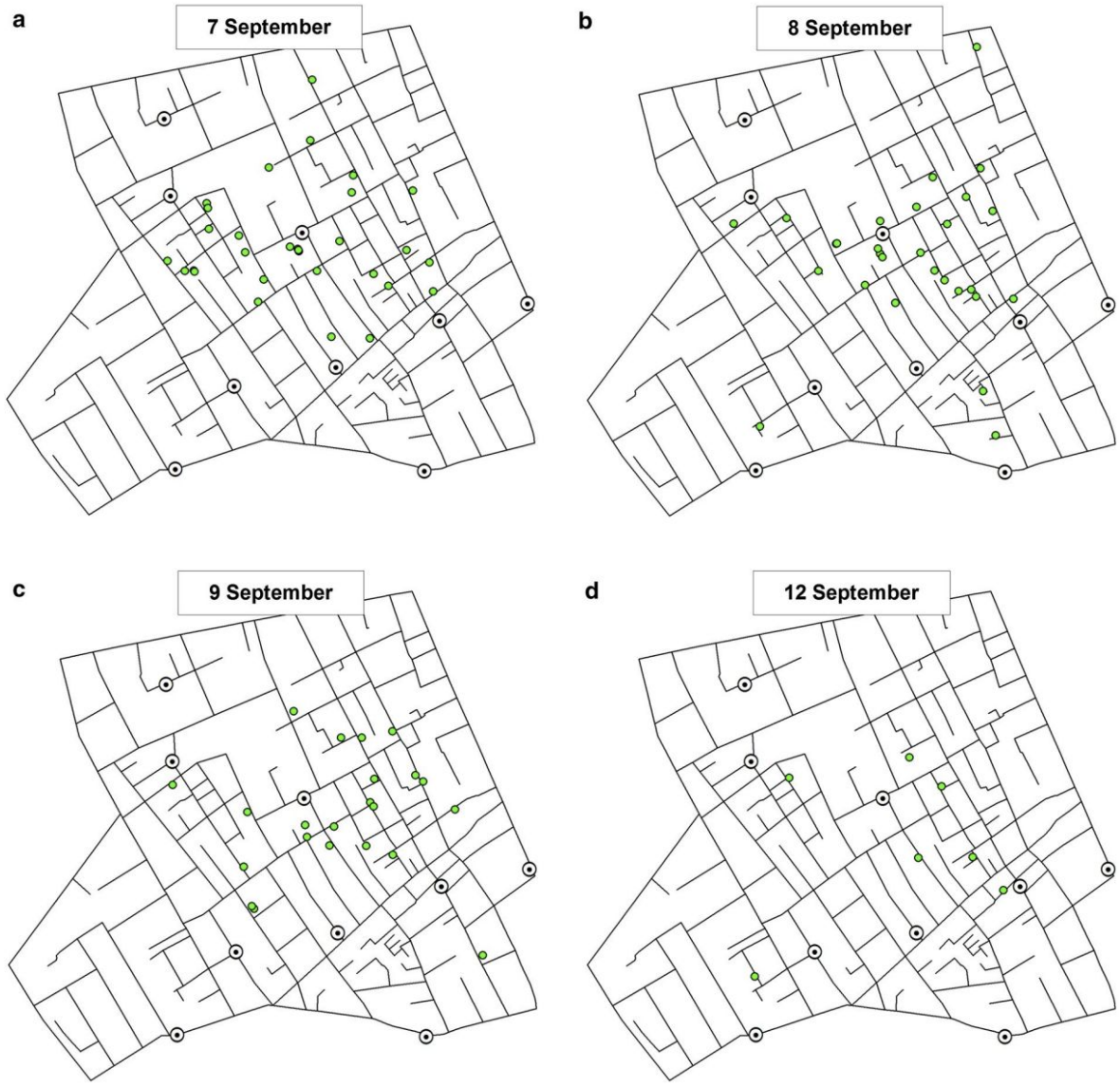

New Death Data
non-matching
matching
Snow's Report only
new
Pumps
Marlborough Mews
Little Marlborough Street / Carnaby St
Broad Street
Dean Street
Rupert Street
Bridle Street
Warwick Street
Vigo Street
Tichborne Street

Figure 3. Shiode et. al (2015) death from cholera recorded in the study area over six-day period

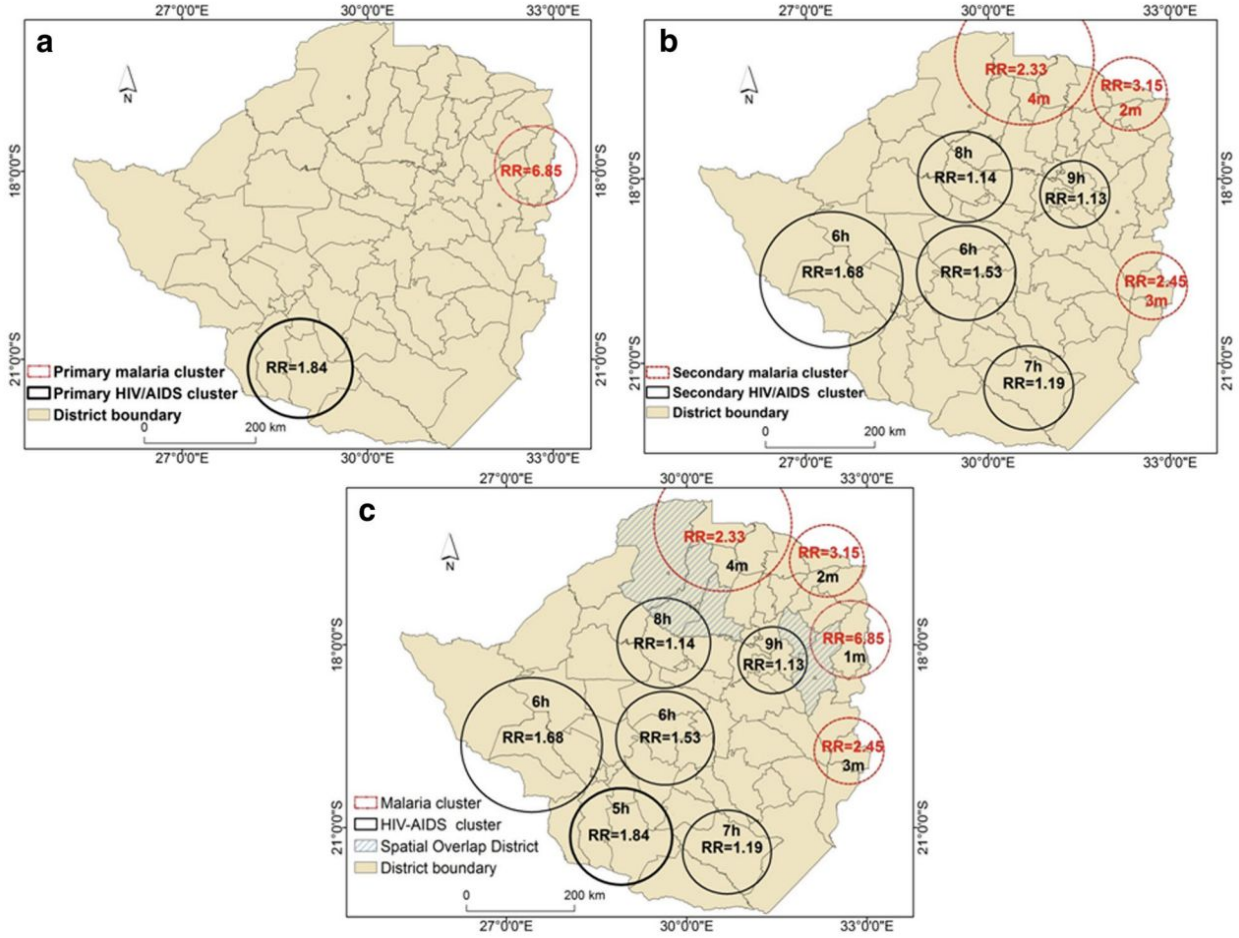


Modern Epidemiology

The field of epidemiology has grown exponentially in the past century. Advances in laboratory testing and mathematical methods for recognizing public health abnormalities have made outbreak containment efforts much more efficient, saving more lives and protecting the wellbeing of people around the world. The computerization of patient intake and disease data has also helped with surveillance measures. Modern spatiotemporal algorithms can find aberrations

in almost any dataset, regardless of nature. This versatility is helpful during viral epidemics, which have a much faster transmission rate than antimicrobial resistance. For example, numerous studies have used scan statistics to track comorbidities with HIV. A 2018 study by Gwitria et. al used case data from Zimbabwe's Ministry of Health and Child Care to examine spatial overlaps in HIV/AIDS and malaria across the country. Their findings noted spatial overlaps in the distribution of secondary clusters of HIV/AIDS and malaria in three northern districts, as well as an overlap in a statistically significant HIV/AIDS secondary cluster with a primary malaria cluster in a Northern and an Eastern district as seen in Figure 4. Another study by Das et. al (2018) used SaTScan's space-time discrete Poisson model on HIV and Hepatitis surveillance data between the years 1980 and 2016 in the District of Columbia. Out of 12,965 diagnosed cases of HIV, 2,316 were co-infected with Hepatitis. Of these co-infected cases, 72.2% were male, 84.6% were of African American descent, and specific populations had greater burden of cases, such as the older age group (60+) and men who have sex with men (MSM). The outcomes of this study identified high risk clusters for comorbidity and the subsequent treatment complications of antiretroviral therapy (ART) by escalating the risk of drug-related hepatotoxicity. The results also contained valuable information about vulnerable persons from the public health perspective. The influence of each variable within an algorithm provided clinicians with a better narrative of resource-limited populations.

Figure 4. Gwitria et. al (2018) Distribution of HIV/AIDS and malaria clusters with relative risk values (RR). Hatching pattern indicates areas of cluster overlap.



In more urgent viral cases, the use of scan statistics can mean life or death. The Ebola virus epidemic ravaged many western African countries from 2013 through 2016, and was the largest Ebola outbreak in history. As reported by the World Health Organization, the explosive epidemic resulted in 28,616 confirmed, probable and suspected cases and 11,310 deaths (www.who.int, 2020). Travel was cited as a major reason for the severity of the epidemic. Previously, Ebola cases were seen in rural, isolated villages, where the afflicted individuals were not likely to travel far distances to receive medical care. In this epidemic, sick people from the rural villages travelled to large urban areas to seek treatment. The continuous travel into densely

populated urban areas and across country borders resulted in a wider spread of the contagion over a matter of days. The unstable nature of an outbreak demands a swift containment response. As shown by Ying-Hen Hsieh (2015), the temporal "waves" of Ebola through Guinea, Liberia and Sierra Leon could be pinpointed with the Richards model. However, this analysis was conducted amidst the crisis, so the temporal waves seen were modeled on reported cases/death, which did not exactly correspond to waves of transmission. As such, Hsieh suggested the resulting clusters should be viewed qualitatively in a relative temporal variation in the reproduction number as opposed to the actual magnitude. Suchar et. al (2018) looked at the same aggregate data from this epidemic retrospectively. From this they hypothesized specific risk factors for the virus, such as having access to bicycles, motorcycles or cars, education and literacy level, number of children living in the home, sharing the toilet with other households and number of sexual partners. Using SaTScan's space-time discrete Poisson model (Monte Carlo Simulations = 99,999) for spatiotemporal analysis and existing demographic data, a wide variety of factors were found to have a significant effect on the Ebola outbreak dynamics including case isolation, safe burials, differences in intervention control and even temperature seasonality.

A relevant pandemic is occurring at the time of our project. The novel coronavirus disease 2019 (CoVid-19) was first identified in Wuhan, China, and as of today, March 18th, 2020 has reached every continent except Antarctica. The symptoms of CoVid-19 include fever, fatigue and dry cough. The latency period of this highly contagious virus ranges from two to fourteen days, and asymptomatic carriers have been identified in the ~197,000 diagnosed cases worldwide (cnn.com, 2020). The virus itself is similar to Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome (SARS). The predictive utility

of scan statistics is paramount to the containment of the disease. Al-Ahmadi, Alamadi and Al-Zahrani (2019) employed SaTScan's spatio-temporal discrete Poisson model (Monte Carlo Simulations = 999) to predict spatiotemporal incidence clustering using country-wide laboratory data of confirmed MERS-CoV cases between June 2012 and March 2019. They found the highest annual risk in ten cities within the Riyadh province of Saudi Arabia between the years 2014 and 2016, and monthly clusters in the cities of Jeddah, Makkah and Taif. These findings provided preliminary risk assessments for further investigation of MERS-CoV clusters, and could easily be applied with greater specificity to the current CoVid-19 crisis worldwide.

While viral infections are volatile and spread rapidly, they are often identified and tended to faster than bacterial infections, specifically in regards to resistance. As mentioned in chapter one, the origin of resistance stems from mutation. Furthermore, poor hygiene and high levels of antimicrobial use have resulted in an increased proportion of resistant strains through selection pressure. As noted by Collignon et. al (2016), the highest use of antimicrobials are seen internationally within food animals, such as in the United States, where nearly 80% of the total volume of antimicrobial use occurs in the animal sector. As such, resistance is a multi-faceted threat that must be addressed on many levels, including animal husbandry. More concerning still, these organisms are capable of acquiring resistance to multiple drugs. Sequencing and genome analysis of such pathogens have revealed acquisition of antibiotic resistance and virulence genes through horizontal gene transfer. For example, as referenced by Juhas (2013), the severity of an *S. aureus* infection is augmented by the staphylococcal cassette chromosome *mec* (SCC*mec*), which harbors resistance genes against tetracycline, kanamycin, streptogramin, bleomycin, lincosamide and tobramycin alongside the more common methicillin and penicillin resistance.

These hyper-virulent “superbugs” can wreak havoc within hospital facilities and communities, and can result in dangerous outbreaks affecting immunocompromised and otherwise-healthy persons alike.

While therapeutic decision-making and clinical judgement should be influenced by laboratory results, a standardized method of classification for resistant microbes should be instilled in microbiology laboratories world-wide for epidemiological and public health purposes. As suggested by Magiorakos et. al (2011), acquired resistance can be generally classified into three categories: multidrug-resistant (MDR), extensively drug-resistant (XDR) and pandrug-resistant (PDR). In order to implement these definitions, ongoing identification of these resistant organisms require that isolates are ideally tested to all antimicrobial agents without selective reporting or suppression of results. This may seem like an excessive practice, but as demonstrated by Semret et. al (2018), research and distant reference laboratories benefit from the etiological abilities offered when differentiating non-malarial febrile illnesses (NMFIs) in low-resource countries. However, rigorous comprehensive testing is treated with caution in developed countries, where the advantages of selective reporting and antimicrobial stewardship reflected an improvement in appropriate prescriptions by general practitioners (Bourdellon et. al, 2017). These contrasting methodologies of AMR diagnostics establish the complexity of understanding antimicrobial surveillance and defining an outbreak.

Introduction to Scan Statistics

Many models exist to define scan statistics for epidemiological data. Spatial statistics detect a local excess of events and test whether this excess is the product of mere chance (Kulldorff, 1997). As discussed earlier in the chapter, these algorithms have a diverse range of

applications within and beyond medicine. One of the early pioneers in this field of mathematics was Joseph I. Naus. Naus built upon earlier work by Berg (1945) and Mack (1948), which determined that given N points independently drawn from the continuous uniform distribution on $(0, 1)$, the probability that the largest number of points clustered within an interval of length p, is greater than or equal to n, defined in notation as $P(n|N;\ p)$. Without diving into the full derivation, Naus continued on to apply his findings to two tangible circumstances:

I. Firstly, he dialed fifteen phones at times distributed at random over a one-minute period, with each dialing period lasting ten seconds. He found that the probability that eight or more phone calls being dialed at the same time could be defined as $'P(8|15;\ 1/6) = 0.037$ in that for some $i = 1, \ldots, 8$, dialing for the $(7 + i)th$ phone call started less than ten seconds after dialing started for the ith phone call, which is a strikingly different outcome from the probability of any one of these intervals contained at least eight points, given six disjointed intervals of ten seconds each, which equated to $6G_b(8|15;\ 1/6) = 0.008$. These findings were based on equation 1.1 (Appendix A). The difference of probabilities exemplifies an aspect of our project, in which a single change in parameter can produce an entirely different result up for interpretation.

II. The second circumstance Nauss applied his findings to was a condition that adheres to the Poisson process and counting the occurrence of an event happening. Here, impulses were generated and received by a counter with a mean rate of λ. In this scenario, the counter registered and defined a cluster whenever n impulses had occurred together in an interval of length less than t. From this, Nauss sought to find the distribution of waiting times until the first registration of the counter. In other words, he wished to quantify the

distribution of the periods of time that occurred after impulse before the recognition of an impulse cluster. As such, the problem was defined as solving for the probability $[F_n(T,\ t|\lambda)]$ that the waiting time until registration of the first cluster was $\leq T$, given that N had a Poisson distribution with a mean of λT and represented the total number of points in $(0,\ T)$. After solving, the probability of waiting time until detection of the first impulse cluster was the sum of conditional probability over the distribution of $N,$ as shown:

$$F_n(T,\ t|\lambda) = \sum_{N=n}^{\infty} P(n|N;\ t/T)\frac{e^{-\lambda T}(\lambda T)^N}{N!}$$

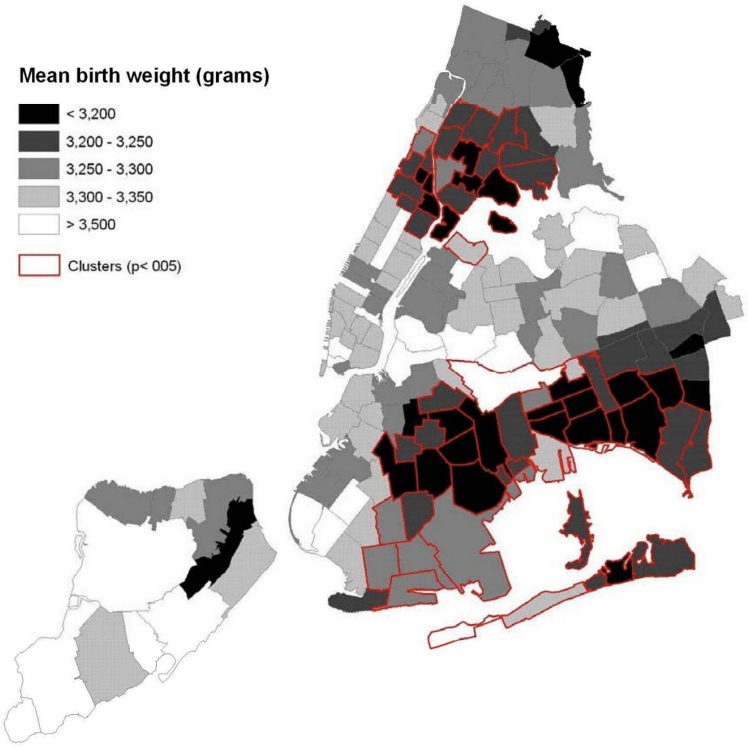
It is important to note that this model came with multiple contingencies. $F_n(T,\ t|\lambda;\ N)$ is the probability $F_n(T,\ t|\lambda)$ conditional over N fixed. $F_n(T,\ t|\lambda)$ is the probability that n (or more) of the N impulses appear in a subinterval of T of length t. The essence of our project's inquiry can be pinpointed here, where Nauss explicitly declares that $F_n(T,\ t|\lambda;\ N) = P(n|N;\ t/T)$ as a result of the continuous uniform distribution being fixed. The continuous uniform distribution is fixed because of the Poisson distribution conditional on the total number N. Even the slightest change in data or misunderstanding of application can create fictional clusters or associations, of which can cause dire consequences if acted upon. What would happen if Nauss had run a different algorithm on these impulse clusters? Would we still receive the same probabilities, and would we be able to credit these clusters to an identifiable event as opposed to random chance?

Space-Time Continuous Uniform Model

In the context of our project, it is important to note that Dr. Martin Kulldorf created a

space-time uniform model specifically for WHONET data interpretation based upon his previous work. As such, there exists no literature on the particular model we will be using for comparisons, but we will consider the continuous uniform distribution model in other studies for foundational understanding. Kulldorf, Hwang and Konty (2009) employed a normal uniform distribution assumption to study low birth rate and early gestation clusters in New York City. In this study, the normal uniform model most accurately fit the data set, since birth weight in New York City is a continuous measure. The study's results were as expected in which the spatial pattern of birth weight simply correlated to demographic and pregnancy-specific demographics. For example, the two statistically significant clusters noted were 61 zip-code regions of eastern Brooklyn and southern Queens, and 29 zip-code regions of northern Manhattan and southern Bronx, $(LLR = 125.8,\ p < 0.001)$ and $(LLR = 62.7,\ p < 0.001),$ respectively (Figure 5). The findings aligned closely with areas already known for increased risk of infant mortality. An evident weakness of the continuous uniform model was seen in results from Staten Island. A non-significant cluster appeared in one singular zip-code from Staten Island of birth weight 60 grams less than average $(LLR = 3.6,\ p < 0.90),$ which could be attributed to chance due to the small number of births within the scan statistic window. This applied to the significant clusters as well, where the high-weight birth areas could potentially be just outside of the window scanned. Moreso, the continuous uniform model operates under the assumption of a normal distribution, and therefore adheres to a likelihood ratio test if the data is truly normal, which is rarely the case. Due to this and the generalized area of clustering as opposed to exact boundaries, the continuous uniform model falls flat in epidemiological analysis of a dynamic population.

Figure 5. Kulldorff, Hwang and Konty (2009) Geographical Distribution of Birth Weight in New York City zip codes in 2004



Space-Time Discrete Poisson Model

The discrete Poisson model is a well-regarded statistical application for assessing the distribution of rare events for count data, such as survival or mortality data, and is closely related to the binomial and Gaussian distributions. Taeger and Kuhnt (2014) provide the example:

> "Of interest are hospital infections on the islands of Laputa and Luggnagg with a conjecture of 4 expected infections per hospital in a year. The null hypothesis is H0 : $\lambda = \lambda 0$, with $\lambda 0 = 4$. The dataset summarizes for how many of the 42 hospitals on the islands the number of infections ranges from zero to six".

To assess Poisson parameters, one can assess between the Poisson's distribution λ versus the λ_0 via Gaussian approximation described by the following test statistic under contingencies described in Appendix A 1.2:

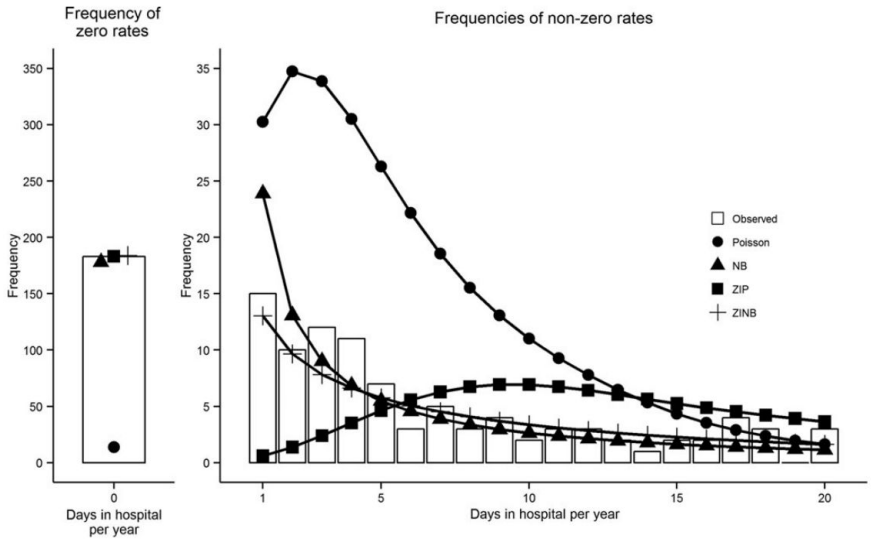
$$Z = \frac{\sum_{i=1}^{n} X_i - n\lambda_0}{\sqrt{n\lambda_0}}$$

Here, the test statistic Z above assumes a standard normal distribution when $n\lambda$ is sufficiently large, meaning the mean and variance follow a Gaussian curve. Additionally, this value can be improved with a continuity correction. Furthermore, an exact test can be executed on the Poisson distribution to determine if the parameter λ differs from a pre-specified λ_0 which is simply the exact version of the Gaussian approximation described above. In the case of two independent Poisson distributions, one can test if the difference between λ_1 and λ_2 is zero through a simple $z-$test (Thode, 1997).

Many aforementioned studies have used SaTScan's space-time discrete Poisson model in an automated fashion to monitor evolving infections seen in microbiology laboratories and hospital data. One such study by Elias et. al (2006) utilized SaTScan's space-time Poisson model to evaluate invasive meningococcal disease (MCD) throughout Germany over a period of 42 months. Different strains were identified by multilocus sequence typing (MLST), and an excess of incidence for identical strains indicated a cluster. Since meningococcal species are universally susceptible to penicillin, this study did not track resistance patterns, but it provides an example of SaTScan's mechanical abilities. The space-time discrete Poisson model assumes the data is Poisson distributed in each demographic region studied, and then the scan statistic is applied to circular windows that originate from user-defined geographic coordinates of increasing size. From this, SaTScan compares observed versus expected cases from inside and outside the window and then decides which of the clusters are the least likely to occur by chance. The location (spatio- component) of each specimen was derived by the postcode and the date of the sample was defined as the time (temporal component) of the illness.

The limitations for the space-time discrete Poisson model assumption are a focus of our project. As reported by Weaver et. al (2015), data often violates the Poisson distribution assumption. When looking at hospitalization records for 313 patients treated with hemodialysis (HD) or peritoneal dialysis (PD), they found that more than half of the patients were never hospitalized, resulting in an excess of zeros. Therefore, when running a Poisson model to estimate the effect of dialysis modality on the rate of hospitalization, they found that HD patients had significantly higher rates of hospitalization ($RR_{HD:PD}$: 1.93; standard 95% CI 1.78-2.10). However, 183 (58%) patients were never hospitalized when compared to 13.6 (4%) expected zero counts from the Poisson analysis, as seen in Figure 6. The predominance of zeros in a data set demonstrates a weakness of the Poisson distribution assumption. This is known as the zero-inflated Poisson (ZIP) distribution, and is a fatal oversight in any analysis.

Figure 6. Weaver et. al (2015) Comparison of Observed Versus Expected Hospitalization Rate Frequencies (Note: NM = Negative Binomial, ZINB - Zero-Inflated Negative Binomial)



Another flaw in the Poisson distribution assumption noted in the same study is the presence of overdispersion. This occurs when the variance of the distribution of the number of

events is not equal to the mean. In other words, a Poisson distribution is not robust to circumstances in which conditional variance is greater than the conditional mean. ZIP is one form of overdispersion in sample data. Another example of overdispersion described by Wang et. al (2012) looks at the number of daily calls within a 15-minute interval from an Israeli call center. Under the assumption that the number of calls from different days are independent, they found sample mean and variances to be 18.66 and 25.95, respectively. With a T_1 *p*-value of 0.07, the presence of overdispersion is evident. Thus, choosing a Poisson distribution assumption for epidemiological data analysis must be conducted alongside rigorous assessment of the data set in question.

Space Time Permutation Model

The space-time permutation model was proposed in 2005 and is the most modern of the three algorithms in our study. Like the Poisson assumption model, the space-time permutation model compiles thousands to millions of overlapping cylinders to create a massive scanning map for any possible clusters (Figure 7). However, the base of these cylinders are often oval, while the Poisson algorithm is circular. As Kulldorff et. al (2005) writes, the novel component of the space-time permutation is the probability model, which does not require any population-at-risk data. Instead, the expected λ must be calculated using only the cases provided. The space-time permutation is ultimately unique in that it bases its predictions off of its many cylindrical tools, defined as:

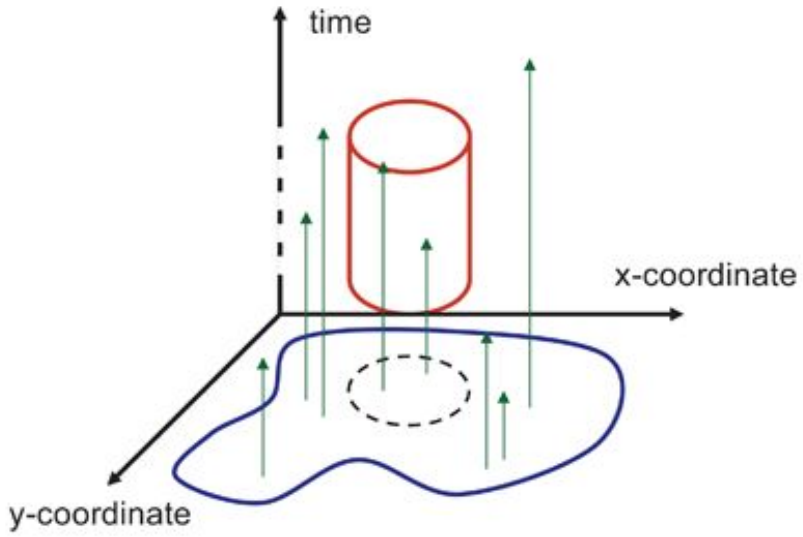
$$\mu_A = \sum_{(z,\, d) \in A} \mu_{zd}$$

In which expected number of cases (μ_A) in hypothetical cylinder *A* equals the sum of the expected

zip-code-days, derived from 1.3 in Appendix A. The "permutation" part of the algorithms title comes in when adjusting for multiple testing. Without any population-at-risk data, adjustments are made through thousands of random permutations using the spatial and temporal attributes of the data set being studied. The most likely statistically significant cluster is simulated for each one of these permutations of the data set, and is then subjected to Monte Carlo hypothesis testing. The end result lists clusters, significant *p*-values, null occurrence rates and outbreak signals.

This algorithm is admittedly robust to missing data and other obstacles that have greater effect on the other two algorithms. However, the interpretation of the space-time permutation demands a greater attention to detail. Kulldorff et al. (2005) used this algorithm to evaluate daily automated analysis of hospital emergency department visits in New York City. The problem arose from numerous false positives and a seemingly hyper-sensitivity to supposed clustering. In contrast to the Poisson-distribution based statistic, the space-time permutation contained excessive random variation in the observed counts that were used to calculate the expected population-at-risk data. So while three highly unusual diarrhea case clusters correctly identified city-wide gastrointestinal outbreaks from rotavirus and norovirus, a number of false alarms were generated and attributed to lack of systemic data quality checks and vigilant analytic adjustments. The space-time permutation offers a compelling new way to detect and manage outbreaks, but requires a more distinct understanding of how the model actually simulates data.

Figure 7. Costa and Kulldorff (2014) Cylindrical Cluster Candidate



Question and Hypothesis

This study investigates the strengths and limitations of three algorithmic models to track and identify emerging antimicrobial resistance and potential outbreaks. The focus of this project is to compare local multiplicities between and within each algorithm. As summarized by Gangnon (2012):

> "Local multiplicity refers to the number of potential clusters that overlap each location. In most, if not all, realistic cluster detection problems, there will be substantial variation in local multiplicity across space. For example, using circular clusters with a fixed maximum geographic radius or fixed maximum population radius, areas with fine geographic resolution will have many more potential clusters than areas with coarse geographic resolution."

Data was processed and analyzed using the free WHONET and SaTScan software under the supervision and guidance of Dr. John Stelling, creator of WHONET software and Dr. Martin Kulldorff, creator of SaTScan software. The aim of this project is to determine the efficacy of three algorithms: the space-time continuous uniform, the space-time discrete Poisson, and the space-time permutation and their ability to detect the same clusters of antimicrobial resistance in retrospective standardized data from Brigham and Women's Hospital. Through various

combinations this project will also investigate the role of three parameters within those algorithms: the maximum spatial size within a population at risk, number of baseline days and maximum temporal size. We hope the results from this study will help researchers understand inaccuracies when describing spatio-temporal trends within the context of antimicrobial resistance and outbreak detection, and perhaps propose more definitive parameters moving forward. Since there have been limited prior studies comparing these specific models within SaTScan, and no prior studies comparing all three, we hypothesize that the three models will not detect an identical number or nature of antimicrobial clusters when applied to the same data set. We further hypothesize the three ancillary variables will also have an effect on detection of significant clusters.

Implication of Research

This investigation will give the field of epidemiology and computational biology a better understanding of somewhat ambiguous algorithm analyses. With more succinct guidelines, the use of spatiotemporal scan statistics will drastically improve outbreak control and prevention. As demonstrated earlier, these models can be applied to nearly any kind of event for aberration detection. While retrospective analyses may be helpful for understanding demographics for potentially susceptible communities, optimized variables and implementation of the correct algorithm can prevent epidemics and predict the nature of disease within different populations. Additionally, the exploration of three variables within the algorithms offers insight into the magnitude of their influence and the way analyses are affected by different parameters. A major limitation and area for future research lies in the validation of the clusters found by each model.

While this project looks at clusters of statistical significance, further studies would benefit from prospective epidemiological validation.

Chapter II

Materials and Methods

This study evaluated three scan statistical algorithms: the space-time continuous uniform, the space-time Poisson distribution and the space-time permutation on seven years of Brigham and Women's microbiology laboratory data. The purpose of these exploratory analyses was to determine the fundamental mechanical differences between each model's outputs. Three additional parameters, baseline days, maximum temporal size, and maximum spatial size in the population at risk, were studied as well to isolate the effect of the algorithms themselves on outbreak detection. Each combination of variables and algorithm was given a code, and geographic clusters were evaluated between codes sharing one or two of the same variable. Each report was run on SaTScan v. 9.6 within WHONET 2019.

Huang et. al (2010) retrospectively reported hospital-acquired infections from Brigham and Women's Hospital (BWH) data between 2002 and 2006 through application of the space-time permutation scan statistic. Hospital-acquired infections were arbitrarily defined by pathogens first isolated >2 days after admission, with clusters only considered significant if the likelihood of occurrence by chance was less than once per 365 days, which were defined by the researchers. When comparing their results to previously-identified cases from routine surveillance, they focused on two types of clusters: first, those with a known epidemiological source of genetically identical strains, and secondly, MRSA and VRE clusters defined as ≥ 3 nosocomial cases in the same ward within a 2 week period. 59 clusters were identified for five

years of data, averaging 12 clusters per year. The two most common types of alerts were for antibiotic resistance profiles. Methicillin-resistant *Staphylococcus aureus* (MRSA) clusters comprised 58% of *S. aureus* alerts and vancomycin-resistant enterococci (VRE) clusters comprised 57% of enterococcal alerts. BWH infection control previously identified two major multidrug resistant clusters of *Acinetobacter* (2004) and *Burkholderia cepacia* (2005), which were also detected by Huang et. al. However, infection control only identified 3 out of the 59 clusters seen by Huang's replication.

Data for Analysis

This project uses routine microbiology laboratory data from 2000 to 2006 for all wards of the Brigham and Women's Hospital, a 750-bed academic medical center located in Boston, Massachusetts. The same data set is the basis of an ongoing study of 80 HCA healthcare hospitals conducted by Channing Laboratory and Department of Medicine at Brigham and Women's Hospital and Harvard Medical School in partnership with the Division of Infectious Diseases and Health Policy Research Institute at the University of California Irvine School of Medicine. We will refer to this as the CLUSTER project.

Analytical Tools

Infection control set rule-based parameters, for which we will define as our "reference" measures. The default number of baseline days = 730, maximum temporal size = 180, and the maximum spatial population = 95. These three parameters are respectively designated by the red, blue and green arrows seen in Figure 5 & 6. These figures refer to "macros". Macros are a WHONET feature for saving the parameters of an analysis. Two types of macros were created: metaward, and metaward-by-resistance. The ward macros are for organisms that are *not*

subjected to routine susceptibility tests, such as bacteroids, fungi, Group A streptococcus and Clostridium difficile, which is commonly associated with hospital outbreaks. These organisms are not routinely tested because they are always susceptible to the usual treatments. The ward-by-resistance macros are for organisms that are subjected to routine susceptibility tests and include all other isolates.

WHONET's software engineer Adam Clark wrote a program to create macros for every permutation of these three variables with the additional values of baseline days = 365, maximum temporal size = 30, 60 or 90 and the maximum spatial population = 50 for each algorithm. A total of 48 macros were created for the three algorithms and the three variables within them (Table 2). Other constant reference parameters include: a minimum of two cases to be considered a cluster, an interval length with first isolate per patient in 365 days and number of Monte Carlo simulations held at 9,999. Lastly, the recurrence interval threshold was kept at 30, of which most statistical texts use as the boundary between small and large samples (Oberle 2015), but we will only be discussing the results from the 365 value. These four variables were held constant throughout the experiment, but future studies may benefit from varying the values.

Hospital surveillance "spatial" locations were defined previously by individual wards and services (such as oncology, surgery etc). Wards and services that share patient care were also accounted for. A "metaward" is a concept relevant to SaTScan's everyday uses, but in our project metawards are groupings of adjacent hospital areas and medical services that would likely be categorized together. For example, Rooms 5A and 5B would be placed in a 5th floor metaward, or IUC-1 and ICU-2 would be grouped into an intensive care metaward. The metaward configuration makes it easier to interpret differences in data between two

fundamentally different spatial locations. The original data file, laboratory configuration and metaward macros were obtained and unaltered in our analysis.

Research Limitations

Since this project is specifically concerned with limitations of these three algorithms, the exploratory data analysis sheds light on the shortcomings of each model by using the same standardized data sets. However, each algorithm has varying levels of sensitivity to sampling volume, variability and distribution. Additionally, data was not adjusted for potential environmental confounders since our focus was the statistical relationship between the three models. SaTScan is one of the most thoroughly evaluated software for epidemiological cluster detection, but a limitation may exist in that this study cannot be directly applied to other government health agencies' epidemiological software. Another possible limitation is homogeneity of the data. The results pertain exclusively to one academic New England hospital, and may not directly translate to other facilities around the world. Huang et. als (2010) earlier study assessed the usefulness of cluster alerts by independent validation from two of the hospital epidemiologists. The epidemiologists were given the characteristics of each cluster (size, type, recurrence interval) and selected one of four responses: ignore, watch and wait, investigate with detailed chart review, or actively intervene with ward-wide cluster precautions. Although they classified 95% of the 59 cluster alerts as useful information, a large majority (73%) of the clusters were deemed as "watch and wait" or "no action necessary". As such, false positives are a real concern when observing seemingly comprehensive results.

A Note on Monte Carlo Simulations (MCS) and Gumbel Locally Adjusted Spatial Scan Statistic (GLASS)

All three models determine statistical significance via Monte Carlo hypothesis testing, where the results of the likelihood function are then compared against a large quantity of random analysis replications of the dataset under the Null hypothesis. In other words, a series of simulations are conducted to evaluate the performance of the proposed bias-corrected estimators (Craiu et al, 2013). It is widely understood that with a higher number of iterations, the greater the precision. SaTScan has three Monte Carlo Simulation iteration options. For this project, we used 9,999 iterations, in accordance with Huang et. al (2010). However, increasing this number to 99,999 iterations may provide results with greater accuracy, and decreasing this number to 999 iterations may provide insight on the significance of Monte Carlo Simulations in relation to each algorithm for cluster detection.

Gangnon (2012) proposed a local multiplicity adjustment using a Gumbel distribution, which he termed Gumbel Locally Adjusted Spatial Scan Statistic (GLASS). As defined earlier, local multiplicity refers to the number of potential clusters that overlap each location. The "local multiplicity problem" was termed by Gangnon (2010), in which differences in the extent of local multiplicities across a study region have an adverse effect on the operating characteristics of a given scan statistic. For example:

> "Under the null hypothesis of constant disease risk, the selection of the most likely cluster is biased in favor of clusters in areas with more extensive local multiplicity (Gangnon and Clayton, 2001; Gangnon and Clayton, 2004). Further, the spatial scan statistic has greater power for single clusters in areas with more extensive local multiplicity than for otherwise identical clusters in areas with lesser local multiplicity (Waller, Hill and Rudd, 2006). Thus, the spatial scan statistic overstates the evidence for clustering in areas with fine geographic resolution and understates the evidence for clustering in areas with coarse geographic resolution."

As such, the GLASS method applies a Gumbel approximation for local scan statistics of each cluster identified. The benefit of this is that GLASS is unaffected by duplicate and overlapping

clusters based on neighborhoods or other spatiotemporal classifications. When Gangnon applied the Gumbel adjustment to a New York Leukemia dataset (Waller et. al, 1994) and a Wisconsin breast cancer dataset (Gangnon and Clayton, 2007), the GLASS statistic outperformed the spatial scan statistic in unbiased cluster detection with reduced edge effects. The GLASS approximation became a permanent function in WHONET due to its clear superiority to the standard scan statistic. As such, our project will be implementing a Gumbel approximation in accordance to the methods used by the CLUSTER project. The application of GLASS may be less effective and potentially adverse in the case of this dataset, which involves a singular hospital, with less distinct geospatial boundaries. To specify cluster's locations, we have applied the same metawards file used by Huang et al. (2010). As mentioned previously, this file further categorizes the locations attached to each isolate. Future studies may benefit from GLASS statistics with our three probability models on a dataset of a larger spatial scope.

Our study is novel, but the results are comparable to similar investigations. Jackson, Baer, Painter & Duchin (2007) conducted a simulation study to compare aberration detection algorithms for syndromic surveillance as well. All simulations and analyses were performed using SAS version 8.2. Using data from Seattle and King county's emergency departments, the study compares the performance of three control-chart-based statistics, two exponential weighted averages and a generalized linear model. The number of baseline days was also considered in the analysis. Variability in daily counts by weekday and month were calculated using a Poisson regression for each syndrome tested. The results indicated that the probability of detecting an outbreak was inversely related to average baseline counts, while our study demonstrated a direct relationship between the two. Across all six algorithms, 310 cluster alerts were observed, and the

probability of detection increased as the size of the outbreak increased. They determined that while significant differences in sensitivity did exist between the models, generally the six models had poor sensitivity, particularly for outbreaks that did not begin with a surge of cases. An important distinction between this study and ours is that SaTScan ran 9,999 iterations of the cylindrical scans, each calculating the probabilities of aberration inside and outside of the oval base, as opposed to a single Poisson distribution calculation for each hypothesis tested. The spatiotemporal nature of our study already indicates a higher caliber for detection of trend anomalies. The sheer volume of these simulations paired with the number of baseline days gave our three models a higher resolution and better resilience towards the stochastic nature of clusters.

In another similar study, Mathes et. al (2017) explored various spatial and temporal scan statistics with emergency department (ED) data from the New York City (NYC) Department of Health and Mental Hygiene (DOHMH). This study evaluated six temporal and four spatial/spatiotemporal detection methods using syndromic surveillance data spiked with simulated injections. Among the spatiotemporal models tested, SaTScan's space-time discrete Poisson and space time permutation were assessed. Overall, they determined that spatial/spatiotemporal methods did not work well in detecting small simulated injects at the zip code level, and sensitivity improved with increased magnitude of injects. This prospective space-time scan study differs from ours by using spiked data and 999 Monte Carlo repetitions. Furthermore, the maximum size of the population at risk was held at 95. Interestingly, although the space-time discrete Poisson scan captured only 3% of all outbreaks, it still found more clusters than the space-time permutation model, which found zero. This finding correlates with

ours, where the Poisson algorithm predominates all other probability models, albeit with ambiguous sensitivity. A possible explanation for this poor performance was discussed by Mathes et. al (2017) in the limitations:

> "[...] We used actual syndromic data for our analytic baseline. Detection of unknown outbreaks or other data anomalies in the baseline data is counted as a false positive, thereby leading to lower specificity and positive predictive values [...] The spatial scan statistic and space-time permutation were the only spatial methods that identified "clusters," i.e., reported signals for a group of neighboring ZIP codes. The other models treated ZIP codes as being independent of each other and did not include any information about how ZIP codes might be spatially related. This information could be included in the models described, however this wasn't attempted due to its complexity."

Further explanations include the computational intensity of running spatial methods and increasing the number of repetitions. The most plausible reason for the incongruity is the size of the data injects. Unlike our dataset, which was actual unaltered syndromic data, the focus of Mathes et. al was detection of small injects of disease. Given that ED data is nuanced, the majority of the probability methods tested were unable to distinguish signals from the rest of the noise. The metawards grouping in our project aimed to neutralize the nuances in the data. Due to these differences, our results may offer a more substantial insight into the mechanism of SaTScans spatiotemporal scans and window variables.

Chapter III

Results

Isolate Summary Categorization

A total number of 18,998 isolates were obtained from 16,549 patients over the seven-year period. The isolates were divided into categories based on their frequency. Isolates with a frequency of ≥ 12 cases per month were deemed "very common", ≥ 4 cases per month as "common", ≥ 1 cases per month as "infrequent" and <1 cases per month as "rare". The most common isolates were from the *Enterococcus species* (46.6 cases/month). These results are expected, as *Enterococcus species* are a common source of urinary tract infections. The least common isolate was *Klebsiella pneumoniae rhinoscleromatis* (0.012 cases/month).

Reference Comparisons

A total of 69 clusters were found between the three algorithms under reference parameters: Baseline days (BLD) = 730, Maximum Spatial Size in Population at Risk (MSPR) = 95 and Maximum Temporal Size (MTS) = 180. Forty-five clusters were detected by only one of the three algorithms, nineteen clusters were found by two of the three algorithms, and five clusters were found by all three algorithms. As demonstrated in Table 4, the space-time discrete Poisson distribution was the most prevalent algorithm for cluster detection by one out of three algorithms (n= 26), with highest recurrence intervals ranging from 370 to 5408. Out of the nineteen clusters detected by two of three algorithms, the space-time discrete Poisson distribution detected all nineteen clusters, and was the most prevalent algorithm once again, with the highest recurrence intervals ranging from 48 to 100,000,000. The second most common model seen was the space-time continuous uniform (n = 12), followed by the space-time

permutation (n = 7). This suggests that the Poisson statistic finds more clusters than the other two algorithms, however, a *statistical* cluster does not necessarily indicate an *infectious disease* cluster. The large volume of clusters found may indicate higher sensitivity to smaller antimicrobial clusters, or may be simple products of spurious correlations. Interestingly, under reference conditions, only the space-time discrete Poisson and space-time continuous uniform algorithms were able to detect the *Burkholderia* clusters, of which were identified as true outbreaks by infection control in March 2005. All three algorithms were able to detect the *Acinetobacter* cluster, identified as a true outbreak by infection control in late October 2004. As described earlier, *Acinetobacter* were categorized as "infrequent" or "rare" depending on the species, meaning < 4 cases per month. *Acinetobacter* clusters 1, 2 and 3 were identified by the Poisson algorithm exclusively, whose signals lasted equal to or longer than three days with the highest recurrence interval at 2133. These clusters, along with similar ones identified by the space-time continuous uniform and space-time permutation models, may warrant further investigation as false positives.

While the clusters found by only one model demonstrate smaller cluster signals, and highlights each model's capacity to identify minor aberrations, clusters found by two models focus on their shortcomings. For instance, cluster 46 (*Achromobacter species*) was detected by the space-time continuous uniform and the space-time discrete Poisson distribution model, but entirely missed by the space-time permutation. This species is considered rare, with a case per month ratio of 0.012, yet two out of three models were still able to notice it. Similarly, clusters 50, 51 and 52 dealt with the rare *Aspergillus species* (case per month = 0.958), which also went undetected by the space-time permutation. In this instance, the space-time discrete Poisson

showed the highest recurrence interval of 100,000,000 out of all clusters detected. The organisms found in cluster 56, *Citrobacter koseri (diversus)* (case per month = 1.321) and cluster 60, *Klebsiella aerogenes* (case per month = 3.798) are both classified as infrequent, and were not detected by the space-time permutation either. As briefly mentioned, the most concerning aberration undetected by the space-time permutation was for cluster 54. The space-time continuous uniform and the space-time discrete Poisson models identified this 2005 outbreak of *Burkholderia*, which was confirmed by hospital infection control.

From the reference assessments, it appears that the space-time discrete Poisson model is the most comprehensive of the three. This may indicate more false positive alerts, but this can be further assessed by looking at the total days and highest recurrence intervals for each signal. Furthermore, the Poisson analyses detected all species and demonstrated some of the highest recurrence intervals, seen in clusters 48, 51, 52, 65 and 67. From the baseline analyses, in which the three models were run on the same dataset with the exact same parameters, the space-time discrete Poisson distribution had the highest local multiplicity.

Parameter Hypothesis Testing

The number of baseline days (BLD) is a user-defined parameter in SaTScan that depends on the amount of baseline historical data entered. This setting can be found in the temporal window tab within SaTScan. Multiple studies have evaluated the role of baseline days within syndromic surveillance systems, but not within the context of our three probability models and two other variables. The number of baseline days is applied by the simulated prospective analysis feature of SaTScan when using historical data (Kulldorff, 2018). The current CLUSTER project has BLD set to 730 days. Our project tested multiple combinations of each

algorithm and variables with the BLD set to 365 days. The macros CU3, PD3 and STP3 were used to isolate the role of baseline days to define a cluster (Table 5). The baseline values for the other two variables were held constant, while the BLD value was changed from 730 days to 365 days. A total of 48 clusters were identified, 21 less than the reference assessment. There were 30 clusters detected by only one algorithm, 18 clusters detected by two algorithms, and zero clusters detected by all three algorithms. The space-time discrete Poisson model detected the highest number of clusters (n = 35), followed by the space-time permutation (n = 16) and the space-time continuous uniform model (n = 15). Some of the clusters missed by the reduced baseline days include organisms such as *Enterobacter faecium, Escheria coli, Serratia marcescens,* etc. While the BLD analysis identified similar temporal clusters to the ones it missed, the considerably lower numbers of clusters overall, and the lack of any clusters identified by all three algorithms demonstrates the importance of baseline days. The analysis also missed the *Achromobacter sp.* (cluster A46) and *Citrobacter koseri (diversus)* (cluster A56) species entirely. Interestingly, the reduced baseline days found two novel clusters that were not identified by the reference parameters (cluster B22: *Haemophilus influenzae* and cluster B48: *Stenotrophomonas sp.*) These clusters may indicate a greater sensitivity when lowering the number of baseline days needed to define a cluster. However, these two clusters may not mean anything substantial if they disappear so easily. As such, we conclude that the number of baseline days plays a significant role in the resolution of each algorithm and in overall cluster detection, but a qualitative analysis of resistance patterns and other cluster details are a consideration for future studies.

The maximum spatial size in the population at risk (MSPR) can be defined by the size of the population and the geographic area of interest. This setting can be found in the spatial

window tab within SaTScan. As mentioned in chapter one, Han et. al (2016) noted that the maximum spatial size is defined at 50% of the total population at risk by default, which generally reduces negative clusters. Studies have selected smaller values in cases of limited data availability, location discontinuity, specific interest in small clusters or rare pathogens, or for limited available resources for intervention (Ma et. al, 2016). The CLUSTER project for which our study is based on currently uses a 95% MSPR, meaning a greater potential number of clusters detected, but with decreased specificity. SaTScan's user manual suggests choosing a high percentage when in doubt, because the scan will detect small and large clusters without pre-selection bias. In our study, we lowered this window down to the standard 50% as experimental values. The macros CU2, PD2 and STP2 were used to isolate the role of maximum spatial size in population at risk values (Table 6). The reference values for the other two variables were held constant, while the MSPR value was changed from 95 to 50. A total of 64 clusters were identified, five less than the reference parameter assessment. There were 31 clusters detected by only one algorithm, 46 clusters detected by two algorithms, and 19 clusters detected by all three algorithms. Again, the space-time discrete Poisson model detected the highest number of clusters (n = 47), followed by the space-time continuous uniform model (n = 26) and the space-time permutation (n = 19). The *Achromobacter species* (cluster A46) from the reference analysis was missed entirely by decreasing the MSPR to 50. The other clusters missed included *Aspergillus sp.* (cluster A52)*, Burkholderia sp.* (cluster A55)*, Klebsiella aerogenes* (cluster A31) and *Staphylococcus aureus* (cluster A43). Besides cluster A46, the decrease in MSPR value was not significant. The clusters missed by this adjustment contained "common" and "very common" isolates, and the reduced window size still correctly identified similar

aberrations during the same period of time for each. From this assessment, we conclude that reducing the maximum spatial size in the population at risk from 95 to 50 does not result in an epidemiologically significant output between algorithms, however, delving into the particulars of these clusters will allow validation of this conclusion.

The maximum temporal size (MTS) is a unit in SaTScan that accompanies the maximum temporal cluster size. In a simulated prospective analysis, the MTS is a flexible window definition that allows evaluation of all temporal windows less than or equal to the specified maximum (Kulldorff, 2018). In other words, with MTS, SaTScan looks for statistically significant clusters in all windows less than or equal to a given period of time. The macros CU4 through CU6, PD4 through PD6 and STP4 through STP6 were used to isolate the role of maximum temporal size. The reference values for the other two variables were held constant, while the MTS values tested were 30, 60, 90. Macros CU7 through CU9, PD7 through PD9, and STP7 through STP9 evaluate the experimental MTS values and MSPR value of 50 while holding BLD at the baseline of 730. Macros CU10 through CU12, PD10 through PD12, and STP10 through STP12 evaluate the experimental MTS values and BLD value 365 while holding MSPR at the baseline of 95. Macros CU13 through CU15, PD13 through PD 15 and STP13 through STP15 evaluate all experimental MTS, BLD and MSPR values. Lastly, the baseline value of 180 was tested with the remaining combinations of BLD and MSPR as seen in macros CU16, PD16 and STP16. A total of 103 clusters were identified, 34 more than the baseline parameter assessment. There were 43 clusters detected by only one algorithm, 42 clusters detected by two algorithms, and 18 clusters detected by all three algorithms. The space-time permutation macros were noticeably less comprehensive than the other two algorithms when adjusting maximum

temporal size, finding only 35 out of the 103 clusters in contrast to 64 (space-time discrete Poisson) and 82 (space-time continuous uniform).

The 43 clusters detected by only one algorithm were analysed first. These clusters were composed of isolates with "common" to "very common" frequencies. The *Acinetobacter sp.* (D1 through D9) and *Burkholderia sp.* (D14 through D18) clusters were of greatest interest since previous outbreaks of these species were confirmed by hospital infection control. For *Acinetobacter sp.,* seven clusters were observed by the reference assessment, and nine were observed when altering the maximum temporal size. Cluster D4 and D9 were detected by only one algorithm: D4 was identified by macros CU10 and CU13 while D9 was identified by PD16. CU10 and CU13 both have baseline days (BLD) = 365 and maximum temporal size (MTS) = 30. However, CU10 has a maximum size of population at risk (MSPR) = 95, while CU13 has an MSPR = 50. Although we previously hypothesized that MSPR does not play a significant role *between* the three algorithms, this finding suggests that MSPR may be more relevant *within* a single algorithm, depending on the other two variables tested. Cluster D9 was identified solely by macro PD16 (BLD = 365, MTS = 180, MSPR = 50). This macro was seen 34 times out of the 64 clusters detected by the space-time discrete Poisson model, making it one of the most comprehensive combinations. Another common pattern was noted for other *Acinetobacter sp.* clusters detected by the space-time continuous uniform and space-time discrete Poisson models. For example, in cluster D6, the macros included CU1, CU4 through CU16 and PD1, PD4 through PD16. The four missing macros (CU2, CU3, PD2, PD3) were from our previous hypotheses testing. All five *Burkholderia sp.* clusters were found by the space-time continuous uniform and space-time discrete Poisson macros, however, the space-time permutation only

managed to identify one of the clusters (D14). The space-time permutation detected this cluster via five macros while the other two algorithms utilized all 16 macros to identify it. From our investigation of the role of maximum temporal size, we can conclude that the space-time permutation does not perform at the same caliber as the space-time continuous uniform or space-time discrete Poisson models within cluster detection, but further investigation of false positives is warranted.

Chapter IV

Discussion

Our results indicated that the space-time continuous uniform, space-time discrete Poisson and space-time permutation algorithms do not detect the identical number or nature of microbial clusters when applied as a simulated prospective space-time scan on the same data set. Furthermore, our study showed that three variables: number of baseline days (BLD), maximum size of population at risk (MSPR) and maximum temporal size (MTS) also have an effect on detection of statistically significant clusters. In particular, when all variables were held constant (BLD = 730, MSPR = 95, MTS = 180), the space-time discrete Poisson probability model detected the highest number of clusters out of the three models tested.

To study the role of each variable we isolated the value by holding all other parameters constant. We found that when the number of baseline days was decreased from 730 to 365, only 69.6% of the original clusters were identified. The space-time discrete Poisson proved to be the most thorough model finding 72.9% of the total clusters. When the maximum size of the population at risk was decreased from 95 to 50, 92.8% of the original clusters were identified. Once again, the space-time discrete Poisson detected the highest number of clusters (73.4%). Lastly, when alternating the maximum temporal size from 180 to 30, 60 and 90 (macros CU4 through CU5, PD4 through PD6, STP4 through STP6), 72 clusters were identified. Including combinations with experimental values BLD = 365 and MSPR = 95, 101 clusters were identified, demonstrating a 46.4% increase in cluster detection compared to the reference analysis. The highest number of clusters was detected by macro PD16 (BLD = 365, MTS = 180,

MSPR = 50), and the space-time discrete Poisson remained the most inclusive algorithm for statistical cluster detection.

From our data, we conclude that the space-time discrete Poisson probability model is the most robust and inclusive analysis within simple statistical cluster detection using a simulated prospective spatial scan within routine microbiology laboratory data, but it's ability to detect *clinically* significant true positives is yet to be determined. Our results also indicate that the number of baseline days and maximum temporal size have a significant impact on cluster detection. The role of maximum size of population at risk remains somewhat inconclusive within the scope of our work.

Future Work

Our results suggest the space-time discrete Poisson model is the ultimate scan statistic: but have we mistaken hypersensitivity for inclusivity? The next step would be to further dissect each cluster and look at the start/end dates, highest recurrence intervals, and the nature of each cluster alert. Since the isolates all come from a singular facility, the metaward categorization (Figures 5 & 6, C:\WHONET\wards6.meta) is pivotal in the spatial areas scanned, and therefore what constitutes a cluster. The most inconclusive area of our project remains in understanding the influence of maximum size of the population at risk (MSPR). We know from many former studies that MSPR can alter the outcomes of probability models, as illustrated by Gleason, Ross and Greeley's (2017) analysis of population-level determinants of legionellosis. Using SaTScan's discrete Poisson probability model, they identified associations of legionellosis with population factors via unadjusted and standardized incident rate methods, and through cluster detection using 1% through 50% MSPR window size reliability score method to minimize

sensitivity. To calculate these reliability scores, SaTScan was run 50 times, increasing the MSPR size by 1% with each run. Each run used a likelihood function to identify the most likely cluster and the secondary statistically significant cluster, with *P*-value determined by 999 Monte Carlo simulations. Their results showed the 1% MSPR value detected 11 statistically significant clusters comprising 93 census tracts, while the 50% MSPR value detected six clusters comprising 259 census tracts. In an effort to understand these numbers, unadjusted odds ratios and 95% confidence intervals of associations with population-level risk factors were applied to the geographic methods of legionellosis. The census tracts classified as high-risk through incidence rate method were then compared throughout the spatial areas. Clearly this is not the ideal analysis. When considering the exclusively "spatial" and "temporal" components of a scan statistic, assessment of categorical variables, whether it be population-level risk factors or four values of an MSPR window, remains an obstacle in data interpretation. Isolation of variables through rigorous testing may alleviate such ambiguity. This is an interesting topic for further research.

Another area of interest would be altering the other variables that were held constant throughout our hypotheses testing, such as interval length from previous isolate, recurrence interval threshold, or minimum number of cases. These constant values were pre-determined by the work we built off of, but could be further investigated with a larger dataset of a greater spatial area. As mentioned in the limitations, increasing the Monte Carlo Simulations to 99,999 and with a Gumbel Locally Adjusted Spatial Scan Statistic (GLASS) could be another future experiment. As described by the literature, we would predict a lower number of clusters, but with a higher resolution and true positive rate.

Ideally, another live prospective study could look at risk values for the cases studied and conduct molecular testing. Retrospectively, the details of each cluster can offer clues to its legitimacy, such as organism type and resistance profile. In past studies, susceptible organisms were overlooked since they were not deemed clinically relevant, while resistant organisms caused concern for transmission. For example, if two patients have the same susceptibility for a sensitive strain, the two cases are probably random. In contrast, if two patients have the same susceptibility of a resistant strain, there is a chance of transmission within the hospital. The next part of this project would involve a retrospective review by infection control personnel to determine what they would consider a true outbreak versus a mathematically interesting (but clinically irrelevant) cluster. Secondly, infection control could comment on whether any "true outbreaks" occurred that would warrant intervention, such as MRSA in the neonatal intensive care unit. These next steps are helpful for refining these algorithms.

Conclusion

Scan statistics play a paramount role in the real-time identification and containment of an outbreak. Retrospective analyses are also crucial as hospital quality control measures and demonstrate the evolution of a pathogen within a community. The application of an algorithm and it's spatiotemporal parameters are pivotal decisions that can drastically alter the analysis outcome. It is the responsibility of the researcher to determine the appropriate probability model and window dimensions according to the nature of the data being studied. With the right execution and interpretation, scan statistics can save many lives and catalyze policy changes for a healthier world.

References

Works Cited

Abellan, J. J., Richardson, S., & Best, N. (2008, April 25). Use of Space–Time Models to Investigate the Stability of Patterns of Disease. *Environmental Health Perspectives, 116*(8), 1111-1119. doi:10.1289/ehp.10814

Alahmadi, S., Al-Zahrani, A., & Al-Ahmadi, K. (2019). Spatiotemporal clustering of Middle East respiratory syndrome coronavirus (MERS-CoV) incidence in Saudi Arabia, 2012–2019. *International Journal of Environmental Research and Public Health, 16*(14), 2520. doi:10.3390/ijerph16142520

Anzia, E. L., & Rabajante, J. F. (2018, September 12). Antibiotic-driven escape of host in a parasite-induced Red Queen dynamics. *Royal Society Open Science, 5*(9), 180693. doi:10.1098/rsos.180693

Bourdellon, L., Thilly, N., Fougnot, S., Pulcini, C., & Henard, S. (2017, January 30). Impact of selective reporting of antibiotic susceptibility test results on the appropriateness of antibiotics chosen by French general practitioners in urinary tract infections: A randomised controlled case-vignette study. *International Journal of Antimicrobial Agents, 50*(2), 258-262. doi:10.1016/j.ijantimicag.2017.01.040

Buckeridge, D. L., Okhmatovskaia, A., Tu, S., O'connor, M., Nyulas, C., & Musen, M. A. (2008, November/December). Understanding Detection Performance in Public Health Surveillance: Modeling Aberrancy-detection Algorithms. *Journal of the American Medical Informatics Association, 15*(6), 760-769. doi:10.1197/jamia.m2799

Collignon, P. C., Conly, J. M., Andremont, A., Mcewen, S. A., & Aidara-Kane, A. (2016, October 20). World Health Organization Ranking of Antimicrobials According to Their Importance in Human Medicine: A Critical Step for Developing Risk Management Strategies to Control Antimicrobial Resistance From Food Animal Production. *Clinical Infectious Diseases, 63*(8), 1087-1093. doi:10.1093/cid/ciw475

Cox, D. R. (1983). Some Remarks on Overdispersion. *Biometrika, 70*(1), 269. doi:10.2307/2335966

Craiu, R. V., Yao, F., & Acar, E. F. (2013). Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics, 7*(0), 2822-2850. doi:10.1214/13-ejs866

Crank, C. W., & O'Driscoll, T. (2015). Vancomycin-resistant enterococcal infections: Epidemiology, clinical manifestations, and optimal management. *Infection and Drug Resistance, 8*, 217-230. doi:10.2147/idr.s54125

Das, S., Opoku, J., Allston, A., & Kharfen, M. (2018). Detecting spatial clusters of hiv and

hepatitis coinfections. *Plos One, 13*(9). doi:10.1371/journal.pone.0203674

Ebola outbreak 2014-2016. (2020, February 17). Retrieved March 29, 2020, from https://www.who.int/csr/disease/ebola/en/

Elias, J., Harmsen, D., Claus, H., Hellenbrand, W., Frosch, M., & Vogel, U. (2006, November). Spatiotemporal Analysis of Invasive Meningococcal Disease, Germany. *Emerging Infectious Diseases, 12*(11), 1689-1695. doi:10.3201/eid1211.060682

Gandy, M. (1999). The Paris sewers and the rationalization of urban space. *Transactions of the Institute of British Geographers, 24*(1), 23-44. doi:10.1111/j.0020-2754.1999.00023.x

Gangnon RE, Clayton MK. A weighted average likelihood ratio test for spatial clustering of disease. Statistics in Medicine. 2001; 20:2977–2987. [PubMed: 11568953]

Gangnon RE, Clayton MK. Likelihood-based tests for detecting spatial clustering of disease. Environmetrics. 2004; 15:797–810.

Gangnon RE, Clayton MK. Cluster detection using Bayes factors from overparameterized models. Environmental and Ecological Statistics. 2007; 14:69–82

Gangnon E. (2009). Local multiplicity adjustments for spatial cluster detection. *Environmental and Ecological Statistics, 17*(1), 55-71. doi:10.1007/s10651-008-0101-0

Gangnon, R. E. (2011). Local multiplicity adjustment for the spatial scan statistic using the gumbel distribution. *Biometrics, 68*(1), 174-182. doi:10.1111/j.1541-0420.2011.01643.x

Gleason, J. A., Ross, K. M., & Greeley, R. D. (2017). Analysis of population-level determinants of legionellosis: Spatial and geovisual methods for enhancing classification of high-risk areas. *International Journal of Health Geographics, 16*(1). doi:10.1186/s12942-017-0118-4

Han, J., Zhu, L., Kulldorff, M., Hostovich, S., Stinchcomb, D. G., Tatalovich, Z., . . . Feuer, E. J. (2016). Using Gini coefficient to determine optimal cluster reporting sizes for spatial scan statistics. *International Journal of Health Geographics, 15*(1). doi:10.1186/s12942-016-0056-6

Hanrahan, L. P., Mirkin, I., Olson, J., Anderson, H. A., & Fiore, B. J. (1990, July). Smrfit: A Statistical Analysis System (Sas) Program For Standardized Mortality Ratio Analyses And Poisson Regression Model Fits In Community Disease Cluster Investigations. *American Journal of Epidemiology, 132*(Supp1), 116-122. doi:10.1093/oxfordjournals.aje.a115772

Hsieh, Y. (2015). Temporal course of 2014 Ebola virus Disease (EVD) outbreak in West Africa

ELUCIDATED Through morbidity and mortality data: A tale of three countries. *Plos One, 10*(11). doi:10.1371/journal.pone.0140810

Ito, T., Okuma, K., Ma, X. X., Yuzawa, H., & Hiramatsu, K. (2003, January 12). Insights on antibiotic resistance of Staphylococcus aureus from its whole genome: Genomic island SCC. *Drug Resistance Updates, 6*(1), 41-52. doi:10.1016/s1368-7646(03)00003-7

Jackson, M. L., Baer, A., Painter, I., & Duchin, J. (2007). A simulation study comparing aberration detection algorithms for syndromic surveillance. *BMC Medical Informatics and Decision Making, 7*(1). doi:10.1186/1472-6947-7-6

Juhas, M. (2013, May 7). Horizontal gene transfer in human pathogens. *Critical Reviews in Microbiology, 41*(1), 101-108. doi:10.3109/1040841x.2013.804031

Kim, S., & Jung, I. (2017, July 28). Optimizing the maximum reported cluster size in the spatial scan statistic for ordinal data. *Plos One, 12*(7). doi:10.1371/journal.pone.0182234

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods, 26*(6), 1481-1496. doi:10.1080/03610929708831995

Kulldorff, M. (2018, March). SaTScan™ user guide for version 9.6. Retrieved April 08, 2020, from https://www.satscan.org/cgi-bin/satscan/register.pl/SaTScan_Users_Guide.pdf?todo=process_userguide_download

Kulldorff, M., & Costa, M. A. (2014). Maximum linkage space-time permutation scan statistics for disease outbreak detection. *International Journal of Health Geographics, 13*(20). doi:10.1186/1476-072x-13-20

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., & Mostashari, F. (2005, March 15). A Space–Time Permutation Scan Statistic for Disease Outbreak Detection. *PLoS Medicine, 2*(3). doi:10.1371/journal.pmed.0020059

Kulldorff, M., Huang, L., & Konty, K. (2009, October 20). A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics, 8*(1), 58-67. doi:10.1186/1476-072x-8-58

Lindsay, J. A. (2013). Hospital-associated MRSA and Antibiotic resistance—What have we learned from genomics? *International Journal of Medical Microbiology, 303*(6-7), 318-323. doi:10.1016/j.ijmm.2013.02.005

Live updates: Coronavirus CASES Top 197,000 globally. (2020, March 18). Retrieved March 19, 2020, from https://www.cnn.com/world/live-news/coronavirus-outbreak-03-18-20-intl-hnk/index.html

Lori, H., William, T., Matthew, S. G., & Tracee, T. (2003, March). The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS). *Journal of Urban Health: Bulletin of the New York Academy of Medicine, 80*(2), 1st ser., 89-95. doi:10.1007/PL00022319

Ma, Y., Yin, F., Zhang, T., Zhou, X. A., & Li, X. (2016). Selection of the maximum spatial cluster size of the spatial scan statistic by using the maximum clustering set-proportion statistic. *Plos One, 11*(1). doi:10.1371/journal.pone.0147918

Magiorakos, A., Srinivasan, A., Carey, R., Carmeli, Y., Falagas, M., Giske, C., . . . Monnet, D. (2012, March). Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: An international expert proposal for interim standard definitions for acquired resistance. *Clinical Microbiology and Infection, 18*(3), 268-281. doi:10.1111/j.1469-0691.2011.03570.x

Marineli, F., Tsoucalas, G., Karamanou, M., & Androutsos, G. (2013). Mary Mallon (1869-1938)
and the history of typhoid fever. *Annals of Gastroenterology, 26*(2), 132-134.

Mathes, R. W., Lall, R., Levin-Rector, A., Sell, J., Paladini, M., Konty, K. J., . . . Weiss, D. (2017). Evaluating and implementing temporal, spatial, and spatio-temporal methods for outbreak detection in a local syndromic surveillance system. *Plos One, 12*(9). doi:10.1371/journal.pone.0184419

Multiple integrals. (2020, March 05). Retrieved March 19, 2020, from https://en.wikipedia.org/wiki/Multiple_integral

Murphy, T. E., Ness, P. H., Araujo, K. L., & Pisani, M. A. (2011, December 24). Bayesian Time-Series Analysis of a Repeated-Measures Poisson Outcome With Excess Zeros. *American Journal of Epidemiology, 174*(11), 1230-1237. doi:10.1093/aje/kwr252

Murwira, A., Mberikunashe, J., Masocha, M., & Gwitira, I. (2018). Spatial overlaps in the distribution of HIV/AIDS and malaria in Zimbabwe. *BMC Infectious Diseases, 18*(598). doi:10.1186/s12879-018-3513-y

National Geographic Society. (2020, March 18). Mapping a London epidemic. Retrieved March 19, 2020, from https://www.nationalgeographic.org/activity/mapping-london-epidemic/

Naus, J. I. (1965, June). The Distribution of the Size of the Maximum Cluster of Points on a Line. *Journal of the American Statistical Association, 60*(310), 532-538. doi:10.2307/2282688

Oberle, W. (2015, July). Monte Carlo Simulations: Number of Iterations and Accuracy. Retrieved March 25, 2020, from https://apps.dtic.mil/dtic/tr/fulltext/u2/a621501.pdf

Rabajante, J. F., Tubay, J. M., Uehara, T., Morita, S., Ebert, D., & Yoshimura, J. (2015, April 22).
Red Queen dynamics in multi-host and multi-parasite interaction systems. *Scientific Reports, 5*(1). doi:10.1038/srep10004

Regunath, H., & Monegro. (2020, January 07). Hospital acquired infections. Retrieved March 19, 2020, from https://www.ncbi.nlm.nih.gov/books/NBK441857/

Semret, M., Ndao, M., Jacobs, J., & Yansouni, C. (2018, April 10). Point-of-care and point-of-'can': Leveraging reference-laboratory capacity for integrated diagnosis of fever syndromes in the tropics. *Clinical Microbiology and Infection, 24*(8), 836-844. doi:10.1016/j.cmi.2018.03.044

Shannon, W. D. (2007). 11 Cluster Analysis. *Handbook of Statistics Epidemiology and Medical Statistics, 27*, 342-366. doi:10.1016/s0169-7161(07)27011-7

Shiode, N., Shiode, S., Rod-Thatcher, E., Rana, S., & Vinten-Johansen, P. (2015). Erratum to: The mortality rates and the Space–time patterns of John SNOW'S cholera epidemic map. *International Journal of Health Geographics, 14*(21). doi:10.1186/s12942-015-0016-6

Suchar, V. A., Aziz, N., Bowe, A., Burke, A., & Wiest, M. M. (2018). An exploration of the spatiotemporal and demographic patterns of Ebola virus disease epidemic in West Africa using open access data sources. *Applied Geography, 90*, 272-281. doi:10.1016/j.apgeog.2017.10.003

Szarka, J. L., Gan, L., & Woodall, W. H. (2011, February 10). Comparison of the early aberration
reporting system (EARS) W2 methods to an adaptive threshold method. *Statistics in Medicine, 30*(5), 489-504. doi:10.1002/sim.3913

Taeger, D., & Kuhnt, S. (2014). Chapter 5: Poisson distribution. In *Statistical hypothesis testing with SAS and R* (pp. 67-75). West Sussex.

Tenover, F. C. (2006). Mechanisms of antimicrobial resistance in bacteria. *American Journal of Infection Control, 34*(5). doi:10.1016/j.ajic.2006.05.219

Thode, H. C. (1997, July). Power and sample size requirements for tests of differences between two Poisson rates. *Journal of the Royal Statistical Society: Series D (The Statistician), 46*(2), 227-230. doi:10.1111/1467-9884.00078

Waller LA, Hill EG, Rudd RA. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. Statistics in Medicine. 2006; 25:853–865. [PubMed: 16453372]

Waller, LA.; Turnbull, BW.; Clark, LC.; Nasca, P. Spatial pattern analyses to detect rare disease clusters. In: Lange, N.; Ryan, L.; Billiard, L., editors. Case Studies in Biometry. New York: John Wiley & Sons; 1994. p. 3-22.

Wang, H., Feng, C., Tu, X., & Kowalski, J. (2012, July). A Comparison of Two Test Statistics for Poisson Overdispersion/Underdispersion. *Applied Mathematics, 03*(07), 795-799. doi:10.4236/am.2012.37118

Warden, C. R. (2008, September 22). Comparison of Poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. *International Journal of Health Geographics, 7*(51). doi:10.1186/1476-072x-7-51

Warden, C., Cudnik, M. T., Sasson, C., Schwartz, G., & Semple, H. (2012, December 27). Poisson Cluster Analysis of Cardiac Arrest Incidence in Columbus, Ohio. *Prehospital Emergency Care, 16*(3), 338-346. doi:10.3109/10903127.2012.664244

Weaver, C. G., Ravani, P., Oliver, M. J., Austin, P. C., & Quinn, R. R. (2015, March 25). Analyzing hospitalization data: Potential limitations of Poisson regression. *Nephrology Dialysis Transplantation, 30*(8), 1244-1249. doi:10.1093/ndt/gfv071

Wei, Y., Huang, Y., Luo, L., Xiao, X., Liu, L., & Yang, Z. (2014, July 9). Rapid increase of scrub typhus: An epidemiology and spatial-temporal cluster analysis in Guangzhou City, Southern China, 2006-2012. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4090214/

Yokoe, D. S., Stelling, J., Placzek, H., Kulldorff, M., Kleinman, K., O'brien, T. F., . . . Huang, S. S. (2010). Automated detection of infectious disease outbreaks in hospitals: A retrospective cohort study. *PLoS Medicine, 7*(2). doi:10.1371/journal.pmed.100023

Appendix A

Figure 5. Baseline Space-Time Permutation Metaward Macro

```
Macro Name = Baseline-BWH-Space-Time Permutation_730_180_95-Metaward
Laboratory = LABUSA.BWX
Study = Cluster alerts (Rows = WARD, Columns = SPEC_DATE by Month)
Cluster alerts:  Analysis method = Space-Time Permutation (Simulated prospective)
// syntax to turn Standard Distribution model on
//Cluster alerts, SaTScan:  ModelType = Uniform
One per patient, unit = By time interval or resistance phenotype, First isolate
One per patient, use time intervals = True
One per patient, interval length with previous isolate = 365
Cluster alerts:  Alert format = Summary
Cluster alerts:  Start date = 2001/01/01
Cluster alerts:  End date = 2006/12/31
Cluster alerts:  Baseline days = 730
Cluster alerts, SaTScan:  MaxTemporalSize = 180
Cluster alerts, SaTScan:  RecurrenceIntervalThreshold = 30
Cluster alerts, SaTScan:  MonteCarloSimulations = 9999
Cluster alerts, SaTScan:  MinimumNumberOfCases = 2
Cluster alerts, SaTScan:  AdditionalInformation = Meta-groups (Automatic, C:\WHONET\wards6.meta)
Cluster alerts, SaTScan:  MaxSpatialSizeInPopulationAtRisk = 95
Data file = Merged 2000-2006.BWX
Output = ACCESS File (Baseline-BWH-Space-Time Permutation_730_180_95_9999-Metaward.mdb)
```

Figure 6. Baseline Space-Time Permutation Metaward by Resistance Macro

```
Macro Name = Baseline-BWH-Space-Time Permutation_730_180_95_9999-WardXAbx
Laboratory = LABUSA.BWX
Study = Cluster alerts (Rows = WARD+PROFILE, Columns = SPEC_DATE by Month)
Options, %RIS and Test measurements:  Combine test methods = True (Etest > MIC > Disk)
Options, Isolate Listing:  Test interpretations = True
Options, Resistance Profiles:  Omit if missing antibiotics = True
One per patient, unit = By time interval or resistance phenotype, First isolate
One per patient, use time intervals = True
One per patient, interval length with previous isolate = 365
Cluster alerts:  Analysis method = Space-Time Permutation (Simulated Prospective)
Cluster alerts:  Alert format = Summary
Cluster alerts:  Start date = 2001/01/01
Cluster alerts:  End date = 2006/12/31
Cluster alerts:  Baseline days = 730
Cluster alerts, SaTScan:  MaxTemporalSize = 180
Cluster alerts, SaTScan:  RecurrenceIntervalThreshold = 30
Cluster alerts, SaTScan:  MonteCarloSimulations = 9999
Cluster alerts, SaTScan:  MinimumNumberOfCases = 2
Cluster alerts, SaTScan:  AdditionalInformation = Meta-groups (Automatic, C:\WHONET\wards6.meta)
Cluster alerts, SaTScan:  MaxSpatialSizeInPopulationAtRisk = 95
Data file = Merged 2000-2006.BWX
Output = ACCESS File (Baseline-BWH-Space-Time Permutation_730_180_95_9999-WardXAbx.mdb)
```

Table 1. Summary of Isolate Frequencies

Frequency	Cases/month	Organism name	# of Isolates	# of Patients
Very common	47.7	*Enterococcus sp.*	4003	3487
	43.7	*Staphylococcus aureus ss. aureus*	3670	3157
	29.8	*Escheria coli*	2503	2198
	19.3	*Klebsiella pneumoniae ss. pneumoniae*	1618	1410
	19.0	*Pseudomonas aeruginosa*	1599	1409
Common	11.3	*Enterobacter cloacae*	946	820
	7.9	*Enterococcus faecium*	660	573
	7.1	*Enterococcus faecalis*	598	516
	6.5	*Serratia marcescens*	548	473
	5.8	*Proteus mirabilis*	486	429
	5.2	*Klebsiella oxytoca*	433	375
	5.0	*Haemophilus influenzae*	419	367
Infrequent	3.8	*Klebsiella aerogenes*	319	280
	3.6	*Citrobacter freundii*	298	261
	3.6	*Acinetobacter baumannii*	298	267
	1.8	*Pseudomonas sp.*	150	132
	1.3	*Citrobacter koseri (diversus)*	111	97
	1.3	*Proteus sp.*	108	90
Rare	1.0	*Aspergillus sp.*	69	69
	0.8	*Burkholderia cepacia*	63	58
	0.4	*Streptococcus, beta-haem. Group A*	32	30
	0.3	*Acinetobacter lwoffii*	27	25
	0.3	*Candida lusitaniae*	23	19
	0.3	*Enterobacter sp.*	21	18
	0.2	*Clostridioides difficile*	16	13
	0.2	*Candida guilliermondii*	13	10
	0.1	*Mycobacterium fortuitum*	7	7
	0.1	*Mucor sp.*	6	6
	0.1	*Mycobacterium chelonae ss. abscessus*	5	4
	0.1	*Burkholderia cepacia complex*	5	5
	0.04	*Klebsiella pneumoniae ss. ozaenae*	3	3
	0.04	*Acinetobacter calcoaceticus*	3	3
	0.02	*Fusarium sp.*	2	2

	0.02 0.02 0.01 0.01	*Rhizopus sp.* *Acinetobacter sp.* *Klebsiella pneumoniae ss. Rhinoscleromatis* *Achromobacter sp.*	2 2 1 1	2 2 1 1

Table 2. Data Analyses Codes (MSPR, BLD, MTS)

Algorithm	Reference	Maximum Spatial Size in Population at Risk (MSPR)	Baseline Days (BLD)	Maximum Temporal Size (MTS) & all other combinations
Continuous Uniform (CU)	CU1: (730, 180, 95)	CU2: (730, 180, 50)	CU3: (365, 180, 95)	CU4: (730, 30, 95) CU5: (730, 60, 95) CU6: (730, 90, 95) CU7: (730, 30, 50) CU8: (730, 60, 50) CU9: (730, 90, 50) CU10: (365, 30, 95) CU11: (365, 60, 95) CU12: (365, 90, 95) CU13: (365, 30, 50) CU14: (365, 60, 50) CU15: (365, 90, 50) CU16: (365, 180, 50)
Poisson Distribution (PD)	PD1: (730, 180, 95)	PD2: (730, 180, 50)	PD3: (365, 180, 95)	PD4: (730, 30, 95) PD5: (730, 60, 95) PD6: (730, 90, 95) PD7: (730, 30, 50) PD8: (730, 60, 50) PD9: (730, 90, 50) PD10: (365, 30, 95) PD11: (365, 60, 95) PD12: (365, 90, 95) PD13: (365, 30, 50) PD14: (365, 60, 50) PD15: (365, 90, 50) PD16: (365, 180, 50)
Space-time Permutation (STP)	STP1: (730, 180, 95)	STP2: (730, 180, 50)	STP3: (365, 180, 95)	STP4: (730, 30, 95) STP5: (730, 60, 95) STP6: (730, 90, 95) STP7: (730, 30, 50) STP8: (730, 60, 50) STP9: (730, 90, 50) STP10: (365, 30, 95) STP11: (365, 60, 95) STP12: (365, 90, 95) STP13: (365, 30, 50) STP14: (365, 60, 50)

				STP15: (365, 90, 50) STP16: (365, 180, 50)

Table 3. Experimental Design Rationale

Logic	Experiment	Control & Why
Baseline Comparison: to evaluate the three algorithms with fixed variables to determine outbreak detection capabilities of the models themselves.	Algorithms are tested on the same data set with identical parameters. Macros tested: CU1: (730, 180, 95) PD1: (730, 180, 95) STP1: (730, 180, 95)	The following parameters are being used by the ongoing CLUSTER project. By maintaining these three variables constant, the output of this analysis will identify the detection capabilities between the three mathematical models: Baseline days (BLD) = 730 Maximum Spatial Size in Population at Risk (MSPR) = 95 Maximum Temporal Size (MTS) = 180
Maximum Spatial Size in Population at Risk (MSPR): to evaluate the role of the maximum spatial size in population at risk within cluster detection by the three models and the relationship with the other two variables.	Values tested are 50 and 95. Macros tested: CU2: (730, 180, 50) PD2: (730, 180, 50) STP2: (730, 180, 50)	Control: Dataset, BLD = 730, MTS = 180 By holding the number of baseline days and maximum temporal size constant, the output of this analysis will identify detection capabilities between and within each algorithm depending on maximum spatial size in the population at risk.
Baseline Days (BLD): to evaluate the role of baseline days required to produce a signal in cluster detection by the three models and the relationship with the other two variables.	Values tested are 365 and 730. Macros tested: CU3: (365, 180, 95) PD3: (365, 180, 95) STP3: (365, 180, 95)	Control: Dataset, MSPR = 95, MTS = 180 By holding the maximum spatial size in the population at risk and the maximum temporal size constant, the output of this analysis will identify detection capabilities between and within each algorithm depending on baseline days.
Maximum Temporal Size (MTS): to evaluate the role of the maximum temporal window in cluster detection by the three models and the relationship with the other two variables.	Values tested are 30, 60, 90 and 180. Macros tested: CU4: (730, 30, 95) CU5: (730, 60, 95) CU6: (730, 90, 95) CU7: (730, 30, 50) CU8: (730, 60, 50) CU9: (730, 90, 50) CU10: (365, 30, 95) CU11: (365, 60, 95) CU12: (365, 90, 95) CU13: (365, 30, 50) CU14: (365, 60, 50) CU15: (365, 90, 50) CU16: (365, 180, 50) PD4: (730, 30, 95)	Control: Dataset, MSPR = 95, BLD = 730 By holding the maximum spatial size in the population at risk and the baseline days constant, the output of this analysis will identify detection capabilities between and within each algorithm depending on maximum temporal size. *Note: some macros have values of MSPR = 50 and BLD = 365. These other combinations were placed in the MTS category for clarity.

	PD5: (730, 60, 95) PD6: (730, 90, 95) PD7: (730, 30, 50) PD8: (730, 60, 50) PD9: (730, 90, 50) PD10: (365, 30, 95) PD11: (365, 60, 95) PD12: (365, 90, 95) PD13: (365, 30, 50) PD14: (365, 60, 50) PD15: (365, 90, 50) PD16: (365, 180, 50) STP4: (730, 30, 95) STP5: (730, 60, 95) STP6: (730, 90, 95) STP7: (730, 30, 50) STP8: (730, 60, 50) STP9: (730, 90, 50) STP10: (365, 30, 95) STP11: (365, 60, 95) STP12: (365, 90, 95) STP13: (365, 30, 50) STP14: (365, 60, 50) STP15: (365, 90, 50) STP16: (365, 180, 50)	

Table 4. Clusters Detected by Models with Reference Values (Macros: CU1, PD1 and STP1)

Cluster #	Organism Name	Total Days	Highest Recurrence Interval	Continuous Uniform (CU1)	Poisson Distribution (PD1)	Space-Time Permutation (STP1)	# of Algorithms
A1	*Acinetobacter sp.*	PD1: 3	PD1: 909		X		1
A2	*Acinetobacter sp.*	PD1: 2	PD1: 667		X		1
A3	*Acinetobacter sp.*	PD1: 3	PD1: 2133		X		1
A4	*Burkholderia sp.*	PD1: 2	PD1: 2300		X		1
A5	*Burkholderia sp.*	PD1: 97	PD1: 2315		X		1
A6	*Escherichia coli*	PD1: 48	PD1: 400		X		1
A7	*Escherichia coli*	STP1: 21	STP1: 417			X	1
A8	*Escherichia coli*	CU1: 159	CU1: 435	X			1
A9	*Escherichia coli*	CU1: 119	CU1: 625	X			1
A10	*Escherichia coli*	CU1: 154	CU1: 2500	X			1
A11	*Escherichia coli*	CU1: 183	CU1: 1429	X			1
A12	*Enterobacter cloacae*	PD1: 1	PD1: 1603		X		1
A13	*Enterobacter cloacae*	PD1: 24	PD1: 500		X		1

A14	*Enterobacter cloacae*	CU1: 170	CU1: 1111	X			1
A15	*Enterococcus faecium*	PD1: 2	PD1: 500		X		1
A16	*Enterococcus faecium*	PD1: 1	PD1: 435		X		1
A17	*Enterococcus faecium*	PD1: 1	PD1: 4259		X		1
A18	*Enterococcus faecium*	CU1: 253	CU1: 476	X			1
A19	*Enterococcus sp.*	PD1: 1	PD1: 1367		X		1
A20	*Enterococcus sp.*	PD1: 3	PD1: 370		X		1
A21	*Enterococcus sp.*	PD1: 1	PD1: 909		X		1
A22	*Enterococcus sp.*	STP1: 7	STP1: 557			X	1
A23	*Enterococcus sp.*	PD1: 1	PD1: 500		X		1
A24	*Haemophilus influenzae*	PD1: 2	PD1: 2358		X		1
A25	*Klebsiella aerogenes*	PD1: 4	PD1: 667		X		1
A26	*Klebsiella aerogenes*	PD1: 4	PD1: 769		X		1
A27	*Klebsiella aerogenes*	CU1: 95	CU1: 1429	X			1
A28	*Klebsiella aerogenes*	PD1: 61	PD1: 588		X		1
A29	*Klebsiella oxytoca*	PD1: 1	PD1: 5408		X		1
A30	*Klebsiella pneumoniae*	STP1: 19	STP1: 714			X	1
A31	*Klebsiella pneumoniae*	CU1: 342	CU1: 2000	X			1
A32	*Proteus mirabilis*	PD1: 1	PD1: 769		X		1
A33	*Proteus mirabilis*	PD1: 2	PD1: 588		X		1
A34	*Proteus mirabilis*	PD1: 2	PD1: 1863		X		1
A35	*Pseudomonas aeruginosa*	CU1: 75	CU1: 714	X			1
A36	*Pseudomonas aeruginosa*	PD1: 1	PD1: 526		X		1
A37	*Serratia marcescens*	CU1: 176	CU1: 476	X			1
A38	*Serratia marcescens*	STP1: 8	STP1: 370			X	1
A39	*Staphylococcus aureus*	STP1: 4	STP1: 1999			X	1
A40	*Staphylococcus aureus*	PD1: 2	PD1: 4829		X		1
A41	*Staphylococcus aureus*	STP1: 2	STP1: 370			X	1
A42	*Staphylococcus aureus*	PD1: 10	PD1: 417		X		1
A43	*Staphylococcus aureus*	STP1: 20	STP1: 144501			X	1
A44	*Staphylococcus aureus*	CU1: 125	CU1: 400	X			1

A45	*Staphylococcus aureus*	CU1: 191	CU1: 5000	X			1
A46	*Achromobacter sp.*	CU1: 68 PD1: 68	CU1: 50 PD1: 48	X	X		2
A47	*Acinetobacter sp.*	PD1: 3 STP1: 3	PD1: 769 STP1: 370		X	X	2
A48	*Acinetobacter sp.*	CU1: 115 PD1: 115	CU1: 10000 PD1: 87046	X	X		2
A49	*Acinetobacter sp.*	PD1: 5 STP1: 5	PD1: 500 STP1: 417		X	X	2
A50	*Aspergillus fumigatus*	CU1: 164 PD1: 164	CU1: 2500 PD1: 400	X	X		2
A51	*Aspergillus fumigatus*	CU1: 145 PD1: 145	CU1: 10000 PD1: 5485075	X	X		2
A52	*Aspergillus sp.*	CU1: 2 PD1: 2	CU1: 10000 PD1: 10000000	X	X		2
A53	*Burkholderia sp.*	CU1: 29 PD1: 29	CU1: 769 PD1: 4431	X	X		2
A54	*Burkholderia sp.*	CU1: 40 PD1: 40	CU1: 909 PD1: 1561	X	X		2
A55	*Burkholderia sp.*	CU1: 33 PD1: 144	CU1: 435 PD1: 5994	X	X		2
A56	*Citrobacter koseri (diversus)*	CU1: 18 PD1: 18	CU1: 2500 PD1: 1000	X	X		2
A57	*Escherichia coli*	PD1: 4 STP1: 4	PD1: 2336 STP1: 1181		X	X	2
A58	*Klebsiella pneumoniae*	PD1: 6 STP1: 6	PD1: 909 STP1: 509		X	X	2
A59	*Klebsiella pneumoniae*	PD1: 1 STP1: 1	PD1: 2448 STP1: 526		X	X	2
A60	*Klebsiella aerogenes*	CU1: 211 PD1: 148	CU1: 10000 PD1: 909	X	X		2
A61	*Serratia marcescens*	PD1: 6 STP1: 6	PD1: 1480 STP1: 722		X	X	2
A62	*Serratia marcescens*	CU1: 159 PD1: 159	CU1: 5000 PD1: 1000	X	X		2
A63	*Staphylococcus aureus*	PD1: 14 STP1: 14	PD1: 5194 STP1: 1844		X	X	2
A64	*Staphylococcus aureus*	CU1: 20 PD1: 20	CU1: 2500 PD1: 14003	X	X		2
A65	*Acinetobacter sp.*	CU1: 12 PD1: 12 STP1:12	CU1: 10000 PD1: 376775 STP1: 476	X	X	X	3
A66	*Klebsiella pneumoniae*	CU1: 166 PD1: 166 STP1: 138	CU1: 10000 PD1: 981 STP1: 556	X	X	X	3

A67	*Staphylococcus aureus*	CU1: 1 PD1: 1 STP1: 1	CU1: 2000 PD1: 234391 STP1: 84236	X	X	X	3
A68	*Staphylococcus aureus*	CU1: 20 PD1: 20 STP1: 20	CU1: 625 PD1: 385 STP1: 556	X	X	X	3
A69	*Stenotrophomonas sp.*	CU1: 20 PD1: 20 STP1: 20	CU1: 625 PD1: 1583 STP1: 747	X	X	X	3

Table 5. Parameter Hypothesis Testing: Baseline Days (BLD)(Macros: CU3, PD3 and STP3)

Cluster #	Organism Name	Total Days	Highest Recurrence Interval	Continuous Uniform (CU3)	Poisson Distribution (PD3)	Space-Time Permutation (STP3)	# of Algorithms
B1	*Acinetobacter sp.*	PD3: 4	PD3: 435		X		1
B2	*Acinetobacter sp.*	PD3: 5	PD3: 455		X		1
B6	*Acinetobacter sp.*	PD3: 3	PD3: 526		X		1
B7	*Acinetobacter sp.*	PD3: 3	PD3: 476		X		1
B11	*Burkholderia sp.*	PD3: 35	PD3: 769		X		1
B14	*Escheria coli*	STP3: 7	STP3: 400			X	1
B15	*Escheria coli*	STP3: 29	STP3: 625			X	1
B17	*Enterobacter cloacae*	STP3: 8	STP3: 651			X	1
B18	*Enterobacter cloacae*	PD3: 1	PD3: 2937		X		1
B19	*Enterococcus faecium*	PD3: 1	PD3: 2934		X		1
B21	*Enterococcus sp.*	PD3: 1	PD3: 455		X		1
B22	*Haemophilus influenzae*	STP3: 14	STP3: 835			X	1
B24	*Klebsiella aerogenes*	STP3:16	STP3: 417			X	1
B25	*Klebsiella oxytoca*	PD3: 1	PD3: 5565		X		1
B26	*Klebsiella pneumoniae*	STP3: 25	STP3: 556			X	1
B27	*Klebsiella pneumoniae*	PD3: 1	PD3: 1408		X		1
B28	*Klebsiella pneumoniae*	PD3: 6	PD3: 385		X		1
B29	*Klebsiella pneumoniae*	CU3: 179	CU3: 1429	X			1
B30	*Klebsiella pneumoniae*	CU3: 172	CU3: 833	X			1
B31	*Klebsiella pneumoniae*	CU3: 180	CU3: 714	X			1
B32	*Klebsiella pneumoniae*	CU3: 166	CU3: 3333	X			1
B33	*Proteus mirabilis*	PD3: 1	PD3: 455		X		1

B34	*Proteus mirabilis*	PD3: 2	PD3: 2222		X		1
B35	*Proteus mirabilis*	PD3: 1	PD3: 833		X		1
B38	*Serratia marcescens*	CU3: 145	CU3: 1000	X			1
B40	*Staphylococcus aureus*	STP3: 7	STP3: 500			X	1
B43	*Staphylococcus aureus*	STP3: 4	STP3: 3812			X	1
B45	*Staphylococcus aureus*	PD3: 20	PD3: 1531		X		1
B46	*Staphylococcus aureus*	PD3: 2	PD3: 1361		X		1
B48	*Stenotrophomonas sp.*	PD3: 1	PD3: 476		X		1
B20	*Enterococcus sp.*	PD3: 3 STP3: 3	PD3: 875 STP3: 625	X	X		2
B3	*Acinetobacter sp.*	CU3: 3 PD3: 3	CU3: 1000 PD3: 625	X	X		2
B4	*Acinetobacter sp.*	CU3: 12 PD3: 12	CU3: 10000 PD3: 809866	X	X		2
B5	*Acinetobacter sp.*	CU3: 115 PD3: 44	CU3: 10000 PD3: 4135	X	X		2
B8	*Aspergillus fumigatus*	CU3: 145 PD3: 55	CU3: 2000 PD3: 79335	X	X		2
B9	*Aspergillus sp.*	CU3: 2 PD3: 2	CU3: 10000 PD3: 2950131	X	X		2
B10	*Burkholderia sp.*	CU3: 144 PD3: 33	CU3: 370 PD3: 1000	X	X		2
B12	*Burkholderia sp.*	CU3: 2 PD3: 2	CU3: 1250 PD3: 14184	X	X		2
B13	*Burkholderia sp.*	CU3: 29 PD3: 29	CU3: 5000 PD3: 14213	X	X		2
B16	*Escheria coli*	PD3: 4 STP3: 4	PD3: 5222 STP3: 2060		X	X	2
B23	*Haemophilus influenzae*	CU3: 2 PD3: 2	CU3: 1250 PD3: 4661	X	X		2
B36	*Pseudomonas aeruginosa*	CU3: 1 PD3: 1	CU3: 714 PD3: 476		X	X	2
B37	*Serratia marcescens*	CU3: 6 PD3: 6	CU3: 1000 PD3: 589		X	X	2
B39	*Staphylococcus aureus*	PD3: 10 STP3: 20	PD3: 526 STP3: 22588		X	X	2
B41	*Staphylococcus aureus*	PD3: 2 STP3: 2	PD3: 500 STP3: 1156		X	X	2
B42	*Staphylococcus aureus*	PD3: 14 STP3: 14	PD3: 7841 STP3: 2983		X	X	2
B44	*Staphylococcus aureus*	PD3: 1 STP3: 1	PD3: 73059 STP3: 22489		X	X	2

B47	*Stenotrophomonas sp.*	PD3: 20 STP3: 20	PD3: 588 STP3: 625		X	X	2

Table 6. Parameter Hypothesis Testing: Maximum Spatial Size in Population at Risk (MSPR)(Macros: CU2, PD2 and STP2)

Cluster #	Organism Name	Total Days	Highest Recurrence Interval	Continuous Uniform (CU2)	Poisson Distribution (PD2)	Space-Time Permutation (STP2)	# of Algorithms
C5	*Acinetobacter sp.*	PD2: 3	PD2: 2133		X		1
C6	*Acinetobacter sp.*	PD2: 2	PD2: 667		X		1
C7	*Acinetobacter sp.*	PD2: 3	PD2: 909		X		1
C10	*Burkholderia sp.*	PD2: 97	PD2: 2315		X		1
C12	*Burkholderia sp.*	PD2: 2	PD2: 2300		X		1
C15	*Escheria coli*	STP2: 21	STP2: 417			X	1
C17	*Escheria coli*	PD2: 48	PD2: 400		X		1
C18	*Escheria coli*	CU2: 154	CU2: 2500	X			1
C19	*Escheria coli*	CU2: 183	CU2: 1429	X			1
C20	*Escheria coli*	CU2: 119	CU2: 625	X			1
C21	*Escheria coli*	CU2: 159	CU2: 435	X			1
C22	*Enterobacter cloacae*	PD2: 24	PD2: 500		X		1
C23	*Enterobacter cloacae*	PD2: 1	PD2: 1603		X		1
C24	*Enterobacter cloacae*	CU2: 170	CU2: 1111	X			1
C25	*Enterococcus faecium*	PD2: 1	PD2: 4259		X		1
C26	*Enterococcus faecium*	PD2: 1	PD2: 435		X		1
C27	*Enterococcus faecium*	PD2: 2	PD2: 500		X		1
C28	*Enterococcus faecium*	CU2: 253	CU2: 476	X			1
C29	*Enterococcus sp.*	STP2: 7	STP2: 556			X	1
C30	*Enterococcus sp.*	PD2: 1	PD2: 500		X		1
C31	*Enterococcus sp.*	PD2: 1	PD2: 909		X		1
C32	*Enterococcus sp.*	PD2: 3	PD2: 370		X		1
C33	*Enterococcus sp.*	PD2: 1	PD2: 1367		X		1
C34	*Haemophilus influenzae*	PD2: 2	PD2: 2358		X		1
C36	*Klebsiella aerogenes*	PD2: 4	PD2: 769		X		1
C37	*Klebsiella aerogenes*	PD2: 4	PD2: 667		X		1

C38	*Klebsiella oxytoca*	PD2: 1	PD2: 5408		X		1
C40	*Klebsiella pneumoniae*	STP2: 19	STP2: 714			X	1
C44	*Klebsiella pneumoniae*	CU2: 342	CU2: 2000	X			1
C45	*Proteus mirabilis*	PD2: 2	PD2: 1863		X		1
C46	*Proteus mirabilis*	PD2: 2	PD2: 588		X		1
C47	*Proteus mirabilis*	PD2: 1	PD2: 769		X		1
C48	*Pseudomonas aeruginosa*	PD2: 1	PD2: 526		X		1
C49	*Pseudomonas aeruginosa*	CU2: 75	CU2: 714	X			1
C51	*Serratia marcescens*	STP2: 8	STP2: 370			X	1
C53	*Serratia marcescens*	CU2: 176	CU2: 476	X			1
C56	*Staphylococcus aureus*	STP2: 2	STP2: 370			X	1
C58	*Staphylococcus aureus*	STP2: 4	STP2: 1999			X	1
C60	*Staphylococcus aureus*	PD2: 10	PD2: 417		X		1
C61	*Staphylococcus aureus*	PD2: 2	PD2: 4829		X		1
C62	*Staphylococcus aureus*	CU2: 125	CU2: 400	X			1
C63	*Staphylococcus aureus*	CU2: 191	CU2: 5000	X			1
C1	*Acinetobacter sp.*	PD2: 5 STP2: 5	PD2: 500 STP2: 417		X	X	2
C3	*Acinetobacter sp.*	PD2: 3 STP2: 3	PD2: 769 STP2: 370		X	X	2
C4	*Acinetobacter sp.*	CU2: 115 PD2: 115	CU2: 10000 PD2: 87046	X	X		2
C8	*Aspergillus fumigatus*	CU2: 145 PD2: 145	CU2: 10000 PD2: 103032	X	X		2
C9	*Aspergillus fumigatus*	CU2: 164 PD2: 164	CU2: 2500 PD2: 400	X	X		2
C11	*Burkholderia sp.*	CU2: 40 PD2: 40	CU2: 909 PD2: 1561	X	X		2
C13	*Burkholderia sp.*	CU2: 29 PD2: 29	CU2: 769 PD2: 4431	X	X		2
C14	*Citrobacter koseri (diversus)*	CU2: 18 PD2: 18	CU2: 2500 PD2: 1000	X	X		2
C16	*Escheria coli*	PD2: 4 STP2: 4	PD2: 2336 STP2: 1181		X	X	2
C35	*Klebsiella aerogenes*	CU2: 95 PD2: 61	CU2: 1429 PD2: 588	X	X		2
C39	*Klebsiella pneumoniae*	CU2: 1	CU2: 2448		X	X	2

		PD2: 1	PD2: 526				
C41	*Klebsiella pneumoniae*	PD2: 6 STP2: 6	PD2: 909 STP2: 509		X	X	2
C43	*Klebsiella pneumoniae*	CU2: 211 PD2: 148	CU2: 10000 PD2: 909	X	X		2
C50	*Serratia marcescens*	PD2: 6 STP2: 6	PD2: 1480 STP2: 722		X	X	2
C52	*Serratia marcescens*	CU2: 159 PD2: 159	CU2: 5000 PD2: 1000	X	X		2
C57	*Staphylococcus aureus*	PD2: 14 STP2: 14	PD2: 5194 STP2: 1844		X	X	2
C2	*Acinetobacter sp.*	CU2: 12 PD2: 12 STP2: 12	CU2: 10000 PD2: 376775 STP2: 476	X	X	X	3
C42	*Klebsiella pneumoniae*	CU2: 166 PD2: 166 STP2: 138	CU2: 10000 PD2: 981 STP2: 556	X	X	X	3
C54	*Staphylococcus aureus*	CU2: 20 PD2: 20 STP2: 20	CU2: 2500 PD2: 14003 STP2: 144501	X	X	X	3
C55	*Staphylococcus aureus*	CU2: 20 PD2: 20 STP2: 20	CU2: 625 PD2: 385 STP2: 556	X	X	X	3
C59	*Staphylococcus aureus*	CU2: 1 PD2: 1 STP2: 1	CU2: 2000 PD2: 234391 STP2: 84236	X	X	X	3
C64	*Stenotrophomonas sp.*	CU2: 20 PD2: 20 STP2: 20	CU2: 625 PD2: 1583 STP2: 747	X	X	X	3

Table 7. Parameter Hypothesis Testing: Maximum Temporal Size (MTS)(Macros: CU4-16, PD4-16 and STP4-16)

Cluster #	Organism Name	Continuous Uniform Macros	Poisson Distribution Macros	Space-Time Permutation Macros	# of Algorithms
D4	*Acinetobacter sp.*	CU10, CU13	N/A	N/A	1
D9	*Acinetobacter sp.*	N/A	PD16	N/A	1
D20	*Escheria coli*	N/A	PD1, PD5, PD6, PD8, PD9	N/A	1
D24	*Escheria coli*	N/A	PD11, PD14	N/A	1
D26	*Escheria coli*	CU1, CU4 through CU9	N/A	N/A	1
D27	*Escheria coli*	CU1, CU4, CU7	N/A	N/A	1
D28	*Escheria coli*	CU1	N/A	N/A	1
D29	*Escheria coli*	CU1	N/A	N/A	1

D35	*Enterobacter cloacae*	CU1	N/A	N/A	1
D36	*Enterococcus faecalis*	CU11, CU12, CU14, CU15	N/A	N/A	1
D38	*Enterococcus faecalis*	N/A	PD1, PD4, PD5, PD7, PD8, PD10, PD13	N/A	1
D40	*Enterococcus faecalis*	CU1, CU6, CU9	N/A	N/A	1
D44	*Enterococcus sp.*	N/A	N/A	STP1, STP4 through STP9	1
D48	*Klebsiella aerogenes*	N/A	N/A	STP11, STP12, STP14, STP15, STP16	1
D51	*Klebsiella sp.*	CU1, CU5, CU6, CU8, CU9	N/A	N/A	1
D52	*Klebsiella sp.*	N/A	PD1, PD4 through PD9	N/A	1
D54	*Klebsiella pneumoniae*	N/A	PD4, PD7	N/A	1
D56	*Klebsiella pneumoniae*	CU16	N/A	N/A	1
D57	*Klebsiella pneumoniae*	CU1, CU4 through CU9, CU11, CU12, CU14, CU15	N/A	N/A	1
D59	*Klebsiella pneumoniae*	CU16	N/A	N/A	1
D60	*Klebsiella pneumoniae*	N/A	N/A	STP10, STP13	1
D61	*Klebsiella pneumoniae*	CU5, CU6, CU8, CU9	N/A	N/A	1
D62	*Klebsiella pneumoniae*	N/A	N/A	STP1, STP5, STP6, STP8, STP9	1
D64	*Klebsiella pneumoniae*	CU16	N/A	N/A	1
D65	*Klebsiella pneumoniae*	N/A	N/A	STP4 through STP16	1
D67	*Klebsiella pneumoniae*	CU6, CU9	N/A	N/A	1
D69	*Klebsiella pneumoniae*	CU1	N/A	N/A	1
D73	*Proteus mirabilis*	N/A	PD16	N/A	1
D74	*Proteus mirabilis*	CU5, CU6, CU8, CU9	N/A	N/A	1
D75	*Pseudomonas aeruginosa*	CU1, CU5, CU6, CU8, CU9	N/A	N/A	1
D76	*Pseudomonas aeruginosa*	N/A	N/A	STP10 through STP15	1
D78	*Serratia marcescens*	CU16	N/A	N/A	1
D79	*Serratia marcescens*	CU1	N/A	N/A	1
D80	*Serratia marcescens*	N/A	N/A	STP1, STP4 through STP9	1
D84	*Staphylococcus aureus*	CU5, CU8	N/A	N/A	1
D88	*Staphylococcus aureus*	CU10 through CU15	N/A	N/A	1
D90	*Staphylococcus aureus*	N/A	N/A	STP4, STP6, STP7, STP9	1
D95	*Staphylococcus aureus*	N/A	N/A	STP1, STP4 through STP16	1

D97	*Staphylococcus aureus*	CU4, CU7	N/A	N/A	1
D99	*Staphylococcus aureus*	CU1	N/A	N/A	1
D100	*Staphylococcus aureus*	CU1	N/A	N/A	1
D102	*Stenotrophomonas sp.*	N/A	PD10, PD13	N/A	1
D103	*Stenotrophomonas sp.*	N/A	PD16	N/A	1
D2	*Acinetobacter sp.*	CU4 through CU11, CU13, CU14	PD1, PD4 through PD16	N/A	2
D3	*Acinetobacter sp.*	CU4, CU5, CU7, CU8	PD1, PD4 through PD9	N/A	2
D6	*Acinetobacter sp.*	CU1, CU4 through 16	PD1, PD4 through PD16	N/A	2
D10	*Aspergillus fumigatus*	CU4, CU7	PD4, PD7	N/A	2
D11	*Aspergillus fumigatus*	CU1, CU5, CU6, CU8, CU9	PD1	N/A	2
D12	*Aspergillus fumigatus*	CU1, CU4 through 16	PD1, PD4 through PD16	N/A	2
D13	*Aspergillus sp.*	CU1, CU4, CU5, CU6, CU10, CU11, CU12	PD1, PD4, PD5, PD6, PD10, PD11, PD12	N/A	2
D15	*Burkholderia sp.*	CU4 through CU16	PD1, PD4 through PD16	N/A	2
D16	*Burkholderia sp.*	CU4, CU5, CU7, CU8	PD1, PD4 through PD16	N/A	2
D17	*Burkholderia sp.*	CU1, CU4, CU5, CU6, CU7, CU8	PD1, PD4 through PD9	N/A	2
D18	*Burkholderia sp.*	CU1, CU4, CU5, CU6, CU10, CU11, CU12	PD1, PD4, PD5, PD6, PD10, PD11, PD12	N/A	2
D19	*Citrobacter koseri (diversus)*	CU1, CU4 through CU15	PD1, PD4 through PD9	N/A	2
D25	*Escheria coli*	CU4, CU7	N/A	STP1, STP4, STP5, STP7, STP8, STP16	2
D30	*Enterobacter cloacae*	CU4, CU7	PD4, PD7	N/A	2
D31	*Enterobacter cloacae*	CU10 through CU14	PD10 through PD16	N/A	2
D32	*Enterobacter cloacae*	CU4, CU5, CU7, CU8	PD1, PD4 through PD9	N/A	2
D33	*Enterobacter cloacae*	CU10, CU13	N/A	STP11, STP12, STP14, STP15, STP16	2
D34	*Enterobacter cloacae*	CU4, CU7	PD1, PD4 through PD9	N/A	2
D37	*Enterococcus faecalis*	CU4, CU7	PD1, PD4 through PD9	N/A	2
D39	*Enterococcus faecalis*	CU4 through CU14	PD1, PD4 through PD16	N/A	2
D41	*Enterococcus sp.*	CU4, CU7	PD1, PD5, PD6, PD8, PD9 through PD16	N/A	2
D43	*Enterococcus sp.*	CU4, CU5, CU7, CU8, CU10, CU13	PD1, PD4 through PD16	N/A	2
D45	*Enterococcus sp.*	CU4, CU7	PD1, PD4 through PD9	N/A	2

D46	*Haemophilus influenzae*	CU10 through CU15	N/A	STP10, STP12, STP13, STP15, STP16	2
D47	*Haemophilus influenzae*	CU4 through CU16	PD1, PD4 through PD16	N/A	2
D49	*Klebsiella aerogenes*	CU4, CU7	PD1, PD4 through PD9, PD11, PD14	N/A	2
D50	*Klebsiella sp.*	CU4 through CU9	PD1, PD4 through PD9, PD11, PD14	N/A	2
D53	*Klebsiella oxytoca*	CU4 through CU15	PD1, PD4 through PD16	N/A	2
D63	*Klebsiella pneumoniae*	CU4 through CU9, CU11, CU14	N/A	STP4, STP7	2
D68	*Klebsiella pneumoniae*	CU1, CU4 through CU9, CU11, CU12, CU14, CU15, CU16	PD1, PD4 through PD9	N/A	2
D70	*Proteus mirabilis*	CU10, CU13	PD4 through PD16	N/A	2
D71	*Proteus mirabilis*	CU4, CU7	PD1, PD4 through PD9	N/A	2
D72	*Proteus mirabilis*	CU4, CU5, CU7, CU8, CU10, CU11, CU13, CU14	PD1, PD4 through PD16	N/A	2
D82	*Serratia marcescens*	CU5, CU6, CU8, CU9, CU11, CU12, CU14, CU15	PD5, PD6, PD8, PD9	N/A	2
D83	*Serratia marcescens*	CU1	PD1	N/A	2
D86	*Staphylococcus aureus*	N/A	PD11, PD14	1, 4 through 16	2
D89	*Staphylococcus aureus*	CU4, CU7	PD1, PD4 through PD16	N/A	2
D91	*Staphylococcus aureus*	N/A	PD11, PD14, PD16	STP1, STP10 through STP16	2
D92	*Staphylococcus aureus*	N/A	PD4, PD5, PD7, PD8	STP4, STP6, STP7, STP9	2
D93	*Staphylococcus aureus*	CU4, CU7, CU10, CU13	PD1, PD4 through PD9, PD11, PD12, PD14, PD16	N/A	2
D96	*Staphylococcus aureus*	CU1, CU4 through CU15	PD1, PD4 through PD16	N/A	2
D98	*Staphylococcus aureus*	CU10, CU13	N/A	STP10 through STP16	2
D1	*Acinetobacter sp.*	CU4 through CU15	PD1, PD4 through PD16	STP1, STP4, STP7	3
D5	*Acinetobacter sp.*	CU1, CU4 through 16	PD1, PD4 through PD16	STP1, STP4 through 9	3
D7	*Acinetobacter sp.*	CU4 through 16	PD1, PD4 through PD16	STP5, STP6, STP8, STP9	3
D8	*Acinetobacter sp.*	C4 through C10, C13	PD4 through PD9, PD1, PD11, PD12, PD14, PD15, PD16	STP1, STP4, STP5, STP7, STP8	3
D14	*Burkholderia sp.*	CU1, CU4 through 16	PD1, PD4 through PD16	STP5, STP6, STP8, STP9	3
D21	*Escheria coli*	CU10, CU11, CU13, CU14	PD10 through PD16	STP10 through STP16	3
D22	*Escheria coli*	CU4 through CU9	PD1, PD4 through PD9	STP1, STP4 through STP9	3
D23	*Escheria coli*	CU10, CU13	PD10 through PD15	STP10 through STP16	3

D42	*Enterococcus sp.*	CU10, CU13	PD1, PD4 through PD9, PD11, PD12, PD14, PD15, PD16	STP11, STP12, STP14, STP15, STP16	3
D55	*Klebsiella pneumoniae*	CU4, CU7	PD1, PD6, PD9	STP1, STP6, STP9	3
D58	*Klebsiella pneumoniae*	CU4 through CU10, CU13	PD1, PD5, PD6, PD8, PD9, PD16	STP1, STP5, STP6, STP8, STP9	3
D66	*Klebsiella pneumoniae*	CU4 through CU10, CU13	PD1, PD4 through PD16	STP1, STP4 through STP11, STP13, STP14	3
D77	*Pseudomonas aeruginosa*	CU10, CU11, CU13, CU14	PD1, PD10 through PD16	STP10 through STP16	3
D81	*Serratia marcescens*	CU4, CU5, CU7, CU8, CU10, CU11, CU13, CU14	PD1, PD4 through PD16	1, 6, 9, 12, 15, 16	3
D85	*Staphylococcus aureus*	CU1, CU4 through CU15	PD1, PD4 through PD16	1, 4 through 16	3
D87	*Staphylococcus aureus*	CU4, CU5, CU7, CU8	PD1, PD4 through PD16	1, 4 through 16	3
D94	*Staphylococcus aureus*	CU1, CU4 through CU14	PD1, PD4 through PD9	STP1, STP4 through STP9	3
D101	*Stenotrophomonas sp.*	CU1, CU4 through CU15	PD1, PD4 through PD16	STP1, STP4 through STP9, STP12, STP15, STP16	3

Mathematical equations:

- 1.1 - Nauss (1965)

  Theorem I: For $P(n|N;\ p)$ defined as the probability of $E(n|N;\ p)$,

  $P(n|N;\ p) = C(n|N;\ p) - R(n|N;\ p),$ for $p \geq 1/2,\ n > (N+1)/2,$

  $= C(n|N;p),$ for $p \leq 1/2,\ n > N/2$

- 1.2 - Taeger & Kuhnt (2014)

  Assumptions:

  Data are measured as counts

  Random variables $X_i,\ i = 1, \ldots, n,$ are Poisson distributed with parameter λ

  Hypothesis:

  $(A)\ H_0 :\ \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$

  $(B)\ H_0 :\ \lambda \leq \lambda_0$ vs. $H_1 : \lambda > \lambda_0$

  $(C)\ H_0 :\ \lambda \geq \lambda_0$ vs. $H_1 : \lambda < \lambda_0$

  Test decision:

  Reject H_0 if for the observed value of z of Z

  $(A)\ z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$

  $(B)\ z > z_{1-\alpha}$

  $(C)\ z < z_{\alpha}$

  P-value:

  $(A)\ p = 2(-|z|)$

  $(B)\ p = 1 - (z)$

  $(C)\ p = (z)$

- 1.3 - Kulldorff et. al (2005)

  Given daily case-counts for zip-code areas, c_{zd} is the observed number of cases in zip-code area *z* during day *d.* The total number of observed cases *(C)* is defined as:

  $$C = \sum_z \sum_d c_{zd}$$

  As such, for each zip-code, the calculated expected number of cases is μ_{zd}, with observed marginals:

  $$\mu zd \frac{1}{C} (\sum_z c_{zd})(\sum_d c_{zd})$$