ABSTRACT

Title of dissertation: MODELING SYNDROMIC SURVEILLANCE AND OUTBREAKS IN SUBPOPULATIONS

Christa D. Pettie, Doctor of Philosophy, 2020

Dissertation directed by: Professor Jeffrey Herrmann
Department of Mechanical Engineering

This research is motivated by the need to assist resource limited communities by enhancing the use of syndromic surveillance (SyS) systems and data. Public health agencies and academic researchers have developed and implemented SyS systems as a pattern recognition tool to detect a potential disease outbreak using pre-diagnostic data. SyS systems collect data from multiple types of sources: absenteeism records, over the counter medicine sales, chief complaints, web queries, and more. It could be expensive, however, to gather data from every available source; subsequently, gathering information about only some subpopulations may be a desirable option. This raises questions about the differences between subpopulation behavior and which

subpopulations' data would give the earliest, most accurate warning of a disease outbreak.

To investigate the feasibility of using subpopulation data, this research will gather and organize SyS data by subpopulation (separated by population characteristics such as age or location) and identify how well the SyS data correlates to the real world disease progression. This research will study SyS how reports of Influenza-like-illness (ILI) in subpopulations represent the disease behavior. The first step of the research process is to understand how SyS is used in environments with varying levels of resources and what gaps are present in SyS modeling techniques. Various modeling techniques and applications are assessed, specifically the Susceptible Infected Recovered "SIR" model and associated modifications of that model. Through data analysis, well correlated subpopulations will be identified and compared to actual disease behavior and SyS data sets. A model referred to as ModSySIR will be presented that uses real world community data ideal for ease of use and implementation in a resource limited community. The highest level research objective is to provide a potential data analysis method and modeling approach to inform decision making for health departments using SyS systems that rely on fewer resources.

MODELING SYNDROMIC SURVEILLANCE AND OUTBREAKS IN SUBPOPULATIONS

by

Christa Daniella Pettie

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:

Professor Jeffrey Herrmann, Chair
Professor Robert Gold
Assistant Professor Allison Reilly
Professor Linda Schmidt
Assistant Professor Monifa Vaughn-Cooke

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Synopsis

Public health agencies and academia have developed and implemented Syndromic Surveillance (SyS) systems as a tool to detect an outbreak using pre-diagnostic data. Pre-diagnostic data types include, but are not limited to, absenteeism reports, over the counter medicine sales, chief complaints, calls to nurse hotlines, Emergency Management Services trip reports, and even web queries. This data is input into a detection algorithm, which is preprogrammed with pattern recognition and associated alert thresholds. The end user is an epidemiologist that translates outputs such as data and alerts from the system into a decision of public health benefit. These systems serve various purposes, such as detection of routine illnesses, like influenza, outbreak early warning, such as a bioterrorism attack, or enhanced situational awareness. SyS systems have been utilized increasingly in public health departments as a supplemental means of surveillance. There are a variety of SyS systems developed by academia, private and public entities; for example, University of Pittsburgh's Real Time Operating Detection System (RODS), Johns Hopkins University's Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE), and Center for Disease Control's (CDC) BioSense (Silva 2013; Lombardo, Burkom, Pavlin., 2004; Stoto, 2005).

One key responsibility of a public health department utilizing SyS systems is to interpret the collected SyS data. A Maryland epidemiologist described one of his main job objectives as "monitoring data streams by looking for any specific occurrences of public health interests to help the community" (J. Russell, personal communication, May 14, 2015). Most notably, epidemiologists need to understand their community, health care seeking behavior of their constituents, seasonal patterns and expectations, and disease evolution (J. Russell, personal communication, May 14, 2015; N. Wang, personal communication, May 13, 2015). CDC identifies the three core functions of public health as follows: assessment, policy development, and community assurance (Centers for Disease Control and Prevention, 2011). SyS data and systems enable these functions specifically, by improving community welfare and increasing data collection as well as automating pattern recognition. (Venkatarao, Patil, Prasad, Anasuya, & Samuel, 2012). Decision makers must consider public health benefit, budget priorities, public acceptability, consistency with values of leadership, and sustainability (Hopkins, 2012). SyS provides a source of early indicators for such public health decisions makers. Based on community needs SyS and data can be used to meet varying needs of resource rich and resource limited communities.

1.2 Problem Statement

Public health entities employ SyS systems to detect disease outbreak, but there are limitations, as the SyS system cannot monitor everyone in the region. SyS systems utilize data collected from predetermined fixed locations; so, there is a chance that the SyS system will not detect the disease outbreak because it is monitoring an unaffected location. There is a higher probability of this occurrence in regions with fewer resources to spend on SyS systems (Venkatarao et al., 2012). Furthermore, there is limited research on the advantages and disadvantages of demographic grouping of SyS data. This research will study how the outbreak detection can be improved through the use of SyS data, epidemic modeling, and subpopulation analysis. Specifically, by studying multiple subpopulations through data analysis, subpopulations experiencing an increase in disease incidence can be characterized. This awareness can lead to increased understanding of SyS systems and enhanced use of SyS data. Identifying and mitigating the inherent risk of SyS systems will help to improve the reliability of the systems (Mostashari and Hartman, 2003). Inherent risk of SyS data and systems emanate in the pre-diagnostic nature of the data and sensitivity of the preprogrammed pattern recognition algorithms. Thus, this study is creating a research opportunity to redefine the allocation of resources and provide a new approach to utilizing SyS data and systems to public health entities in resource limited areas. Answering the following research questions will address each aspect of the problem statement.

1.3 Research Questions

1) Which subpopulation's behavior best represents the overall encounters of the SyS data set?

Task #1: Compare the SyS ILI reporting activity of age groups and set geographic locations to the complete (total population) SyS data set. Determine which age group and which location is most similar to the complete data set.

Task #2: Identify peaks in SyS ILI activity and compare the ILI reporting activity of age groups and boroughs of the five weeks leading to the peaks.

2) Using the most representative age group and location (found from question #1), can analysis determine which subpopulation is more likely to miss an increase in overall ILI activity?

Task: Compare SyS ILI summary statistics & reporting activity and activity of the best age subpopulation & best borough to the complete (total population) SyS data set.

3) Which, if any, subpopulation(s) show leading indicators of an impending outbreak? Which demographic category is better to use (an age group or location) to represent the population?

Task: Determine the correlation of ILI SyS reporting activity of subpopulations to confirmed influenza reports and complete ILI SyS data set. Focus on preceding weeks leading to peaks in flu season and consider correlation with a one week and two week lag.

4) Can a model utilizing subpopulation data and/or behavior provide representative and predictive information for the entire population?

Task: Identify and substantiate an epidemic model that can use subpopulation behavior to provide insight to public health decision makers about a potential outbreak.

1.4 Overview

Chapter 2 will present a literature review of SyS systems, health care related studies of resource limited areas, and epidemic modeling techniques. The literature review will discuss SyS systems in greater detail than provided in the brief introduction and explore gaps and limitations in previous epidemic modeling efforts. Chapter 3 will detail the research approach that will be used to answer the research questions. Chapter 4 presents results of data analysis and application of ModSySIR. Chapter 5 discusses the results and implementation of the approach for a public health entity. Finally, Chapter 6 provides a summary, contributions, research impacts, and limitations.

Chapter 2: Literature Review

This chapter will expand on syndromic surveillance capabilities and system evaluations. Epidemic and analogous modeling efforts, specifically the historical applications and modifications of the Susceptible Infected Recovered (SIR) model, will be evaluated to identify a predictive epidemic model suitable for this research. Health care seeking behavior of a community members impact a public health department's decision making and response to public health matters; therefore, community behaviors and outbreak planning, response, and public health policy will be described. Many models are inclusive of transmission and individual contact parameters to show how social contacts affect disease spread. Finally, an implementation guide for this approach will be presented based on an opportunity to build on gaps in other research.

2.1 Syndromic Surveillance Systems

SyS systems serve various purposes such as situational awareness, outbreak detection, and data analysis. A CDC working group (2004) described SyS as "an investigational approach where health department staff, assisted by automated data acquisition and generation of statistical alerts, monitor disease indicators in real-time or near real-time to detect outbreaks of disease earlier than would otherwise be possible with traditional surveillance". SyS is a complement to traditional means of public health surveillance such as laboratory testing or imaging reports. Researchers have used statistical and content based evaluation techniques to assess and compare performance and functional features among these SyS systems and algorithms.

One branch of SyS research analyzes these outbreak detection algorithm variations and performance. Such algorithms are in place with the goal of detecting unusual statistical behavior (Lombardo & Buckeridge, 2007). Algorithm performance can be assessed through the use of control charts. For example, control charts such as cumulative sum, exponentially weighted moving average, and seven day mean, show variations in data. Key parameters such as sensitivity and specificity can feed into performance assessments. Simulations are useful in testing algorithm performance and accuracy and may or may not utilize real outbreak data and event statistics. Some countries have abundant monetary and human resources and a strong public health infrastructure such as France (Josseran et al., 2010), England and Wales (Doroshenko et al., 2005); therefore, they can develop more comprehensive SyS systems and evaluate their performance (Jefferson et al., 2008).

Evaluations of SyS systems in low resource and developing countries can vary greatly based on available resources, political support, and public health infrastructure. May, Katz, Test, and Baker (2011) conducted an evaluation of fourteen resource poor areas utilizing SyS systems. This overview concluded that personnel and training challenges of these areas must be properly managed and identified. Additionally, the collection and transmission of SyS data can prove to be difficult for rural and low resource areas (Soto et al., 2008; Jia and Mohamed, 2015). Understanding the system and public health related limitations, such as technology and availability of health care, of these systems in low resource areas, can help to identify short term solutions or potential improvements for outbreak detection. Capabilities vary depending on the community and its available resources. Additional research could help to better 1) understand public health infrastructure and resource allocation among economically and culturally diverse communities and 2) how SyS can appropriately supplement surveillance in each community based on resource availability.

The user perspective is invaluable in understanding the role SyS plays in the public health surveillance for their health department. Through a series of interviews with US-based epidemiologist, benefits and limitations of SyS systems were explored. Interview questions addressed topics such as data validity, SyS benefits, and community needs. These epidemiologists were from health departments in Philadelphia (Pennsylvania), Montgomery County (Maryland), Houston (Texas), Washington, D.C., and South Dakota. Common insights are listed below:

- Based on professional experience with SyS systems, each epidemiologist set their own threshold for reporting based on the number of SyS alerts protocol for data analysis in which would lead to launching an investigation.
- Data from other public health monitoring efforts and surveillance systems compliment SyS systems.
- De-identified information (SyS data characteristics) from hospitals can make follow-ups and investigations more difficult.
- Situational awareness and event based surveillance are additional benefits of SyS systems.
- Health departments in varying communities, lack resources (specifically people and time) to analyze SyS data as they would like to.
- SyS systems are missing alternate data sources such as urgent care treatment facilities and social media. These would be a great supplement for added awareness.

2.1.1 Data Sets

SyS systems are informed by one or more pre-diagnostic data sources, but epidemic models incorporate many different data types and combinations of sources (Azarian, Winn, Zaheer, Buehler, & Hopkins (2009). Depending on the model, data can be analyzed in real-time or retrospectively. Similarly, social media analytics research has used trends in user conversations to characterize behavior and sentiment. Achrekar, Gandhe, Lazarus, Yu, & Liu (2011) utilized an autoregressive predictive model to determine the correlation of flu related tweets and CDC ILI cases. Google Flu Trends, source of data based on user initiated search queries, is restricted to unsubstantiated information collected online rather than at the point of care.

Analysis can lead to diverse statistical calculations and categorization of SyS data. Age group separation was utilized to analyze and organize flu data (Fleming, Zambon, and Bartelds, 2000). The goal was to identify which group (Age groups: 0-4, 5-14, 15-44, 45-64, 65+) if any peaked before or after all age groups in the population. For one data set of the 1989-1990 flu season, age groups 0– 4 and 5–14 years showed an initial increase of incidence and older age groups demonstrated a one week lag for increase. Cooper, Verlander, Elliot, Joseph, & Smith (2007) determined thresholds, peak weeks, and compared time lag indicators for each age group. SyS and laboratory data was used for early warnings for the start of flu season based on thresholds for each age group. Neither of these efforts produced a predictive model.

Historical data sets, which can also be considered training data, help to set a baseline for deviations and variation indicative of an increase in reports or outbreaks. Retrospective analysis models are trained or baselined on historical data sets which may or may not be indicative of the potential epidemic. Hall, Gani, Hughes, & Leach (2007) utilized data from three influenza pandemics from 1918, 1957 and 1968 to train and validate their model predictions of timing, amplitude, and duration of maximum prevalence of given pandemic wave. Both training and real time data could be from sources such as emergency departments (Josseran et al., 2010), general practitioners (Hall et al., 2007), over-the-counter medicine sales, or even simulation (Valle, Clark and Zhao, 2011). Understanding characteristics of the data set is an important step in having a model that represents a community accurately.

2.2 Public Health

2.2.1 Influenza

Each flu season brings a potential evolution of each previous strain of influenza and public health departments must prepare for the effects of a potential outbreak in their community. There is abundant research about the disease behavior and progression of influenza. Thompson, Comanor, and Shay (2006) estimated the disease burden to determine relative risk in subpopulation and assist with planning for flu season such as vaccination strategy. Another effort studied six pre-pandemic flu seasons and identified the relative risk with this calculation: Relative Risk = (Proportion of Cases before the peak)/ (Proportion of Cases after the peak) (Worby et al., 2015). Cox, Brammer, and Regnery (1994) assessed global laboratory surveillance of emergence of pandemic strains of influenza. This has led to many models and simulations of the disease spread and quantification of identified parameters. While influenza is widely studied, there are still gaps in research. There is a much greater emphasis on resource rich areas such as United States and European Counties. Studies of subpopulation identification and characterization considering transmission in low resource communities is limited. Krumkamp et al. (2011) evaluated an availability model to identify shortfalls in health care resource allocation considering the H1N1 outbreak in Thailand. The study concluded that developing countries need an appropriate way to deal with sparse data collection. Characterizing demographic trends of influenza provides a different type of insight about potential effects of seasonal and pandemic influenza (Thompson et al., 2006). Health departments need to understand the disease behavior and how that will affect the behavior of individuals within the community.

2.2.2 Health Care Seeking Behavior

Understanding the health care seeking behavior (HCSB) patterns in individuals, populations, and areas (i.e. urban vs. rural) can inform epidemic models. Various epidemic models represent different aspects of disease behavior by introducing and modifying parameters and transition rates. By examining both HCSB and modeling, communities can gain insights such as the following:

(1) Illnesses likely to be (or not to be) reported

(2) Length of time one will wait to seeking treatment

HCSB surveys supplement research efforts spanning many communities and illnesses. Using HCSB surveys, researchers in Nigeria found that tuberculosis treatment delays were high and recommended steps to reduce that delay (Ukwaja, Alobu, Nweke, & Onyenwe, 2013). Additionally, Jacobsen et al. (1993) used a male based population study to predict a man's decision to seek medical care based on the severity of his symptoms. This survey identified the likelihood of seeking care for each age group compared to another and what types of obstacles prevent seeking care. Community specific results of HCSB surveys can provide key to insights to inform representative epidemic modeling. Cultural implications can also provide insight about a community. Read et al. (2014) explored the concept of cultural context, finding that number of contacts were the same for individuals of the same culture in both rural and urban areas. The only major differences were seen in areas with high population density. Section 2.3.3 further explores social contacts about how it affects transmission in different settings.

One goal of this research is to characterize various subgroups through in depth analysis of a population. Collected SyS data provides the means to better understand health care seeking behavior of a population; however, it is not enough to presuppose the rationale behind health care decisions of individuals based on SyS data alone. There is a personal decision process that leads an individual to seek care based on factors such as level of illness, societal norms, family, finances, beliefs, and various demographics (Andersen and Newman, 2005; Andersen, 2008). Health surveys and questionnaires have been administered by researchers and health officials to understand how and when community members access health information and care. Differences in these behaviors are observed and categorized based on demographics such as ethnicity, age, and community (van der Hoeven et al., 2012; Lee, Boden-Albala, Larson, Wilcox, & Bakken, 2014; Metzger, Hajat, Crawford, & Mostashari, 2004). Investigating these differences can provide indications of patient and behavioral risk factors (Donaldson et al., 2009; Metzger et al., 2004). Detailed research endeavors are devoted solely to gathering this type of behavioral data. While this analysis effort will not rely on the surveyed community and interview results, it will reference other studies of NYC communities that have relevant population and behavioral data (Lee et al., 2014). Understanding community behavior informs decision making, disease response, and prevention efforts of health agencies.

2.2.3 Public Health Policy and Response

One goal of this research is to provide public health entities with an approach to optimally use their resources through the use of modeling. Many communities have response plans in place for outbreaks and even prior to that, there are vaccination strategies in play. Researchers have explored such plans, interventions, and responses of public health agencies. For example, Goeyvaerts et al. (2015) evaluated the impact of age group specific vaccinations. Uscher Pines, Duggan, Garoon, Karron, & Faden (2007) reviewed national pandemic preparedness plans from high, middle, and low income countries. Plans showed lack of involvement of disadvantaged groups in pandemic planning. These groups need to be identified and engaged during planning; furthermore, special needs of such groups need to be addressed for an appropriate response to an influenza pandemic. Specifically, there is a need to optimize resource use and response plans in advance of influenza outbreaks. Intervention strategies of influenza outbreaks needs to be customized to the dynamics of the infected population (Lee, Kim, and Kwon, 2013). Decision support systems can be utilized from Machine Learning (ML) efforts and assist with monitoring and forecasting to determine appropriate response strategies (Brownstein et al., 2017). Effective surveillance strategies should be developed prior to an epidemic. (Thompson et al., 2006). This topic of the utility and state of ML in consideration of SyS will be discussed later in this chapter.

2.3 Modeling

For epidemic simulation, both retrospective and predictive models have evolved in complexity along with public health surveillance efforts. Models have been developed and evaluated based on parameters and characteristics such as data type, disease behavior, community, social contacts, and surveillance type. This section will present various models that employ at least one of the following: SyS data, influenza, epidemic forecasting/predictive capabilities, social contact matrix, or subpopulation stratification. A brief evaluation of some useful elements identified within these models will be presented, then the most relevant information will be detailed. Many epidemic models have evolved from the Susceptible-Infected-Recovered (SIR) model, which will be explored further in the next section.

2.3.1 Epidemic Modeling

The SIR model was first introduced by Kermack and McKendrick (1927). SIR is a compartment model and assumes a constant total population. There are both a continuous and discrete version of the SIR model.

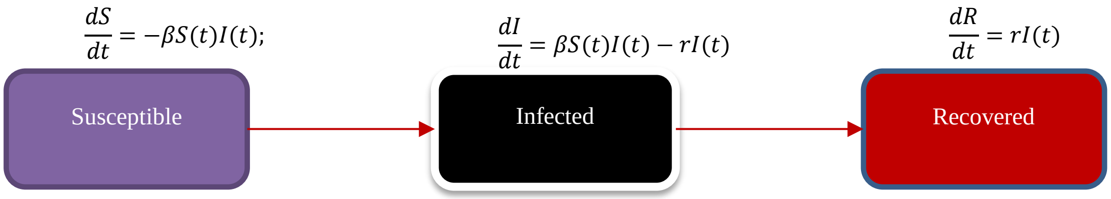

$\frac{dS}{dt} = -\beta S(t)I(t);$
$\frac{dI}{dt} = \beta S(t)I(t) - rI(t)$
$\frac{dR}{dt} = rI(t)$
Susceptible
Infected
Recovered

Figure 1 A graphic showing the relationships and transition rates of the SIR compartments

As shown in Figure 1, individuals within the population will flow from the Susceptible to Infected compartments with a certain transition rate. Then the infected move to Recovered at another transition rate. While individuals may flow between compartments, the overall population (N) will remain unchanged. Let S(t), I(t), and R(t) be the number of susceptible individuals, infected individuals, and recovered individuals, respectively, at time t; therefore, N=S(t)+I(t)+R(t). Usually, N=1 depending on the initially values. In the case of N=1, the initial values would be considered as a fraction of the population and if N is any other number such as 100 or the actual number of individuals in the population the initial values would still be a portion of the population but scaled from 0-100 (or 0 to the population count) and the proportion of contacts from infected individuals that result in infection of susceptible individuals is represented as $\frac{\beta}{N}S$. Important assumptions or definitions are listed below:

1) Susceptible individuals are not yet infected, but are able to be.
2) Infected individuals have the disease and are capable of spreading the disease.
3) Recovered individuals had the disease and are no longer able to spread the disease or become infected again.
4) Initially, R(0) = 0.

Let β be the infection rate and r the recovery rate per unit time. The SIR differential equations are as follows:

$$\frac{dS}{dt} = -\beta S(t)I(t);\ \frac{dI}{dt} = \beta S(t)I(t) - rI(t);\ \frac{dR}{dt} = rI(t)$$

The transition calculations βSI and rI are the rate of new infections and of recovery, respectively. The Susceptible, Infected, and Recovered populations change in accordance to transition rates. These two transition rates are: 1) Susceptible to Infected, the probability of a contact between a susceptible individual and an infectious individual, and 2) Infected to Recovered, one divided by the mean time to recover. These rates depend upon the disease behavior and the characteristics of the individuals or community.

The reproduction rate, R_o, accounts for the average number of people potentially infected in an entirely susceptible population without consideration of an effort to controlling the infection (i.e. no intervention from a public health department or change of behavior of the susceptible population). Alternatively, R_o is also equivalent to the rate of infection divided by the rate of recovery (Beckley et al., 2013). That is, $R_o = \beta/r$. When this is greater than 1, a disease spreads; but when it is less than 1, the disease subsides.

Scholars have made many modifications the SIR model to account for varying disease and population behaviors. Enhancements to the SIR model have combined epidemic modeling with a wide variety of topics including vaccination strategies, global stability, and even consider time delay of certain disease behaviors (Beretta & Takeuchi, 1995; d'Onofrio, 2005; McCluskey, 2010). For example, the SIR model does not include mortality, which is important in producing a representative and realistic epidemic model; however, later modifications of the SIR model account for this by including both death rate and birth rate (Beckley et al., 2013).

Realistically, the probability of disease infection will vary based on location, season, population, and other factors. Variable contact rates and additional compartments are important to consider in representative epidemic modeling. One version of the SIR model is Susceptible-Infected-Susceptible (SIS), in which an individual does not gain immunity once they overcome the infected phase. Hence, one can reacquire the disease, so there is no “Recovered” state. Wierman and Marchette (2004) utilized SIS to model a computer virus spread. Ball, Sirl & Trampan (2010) studied stochastic and deterministic models of SIS population in search of a good indicator for the behavior of households during epidemics. The Susceptible-Exposed-Infected-Recovered (SEIR) set of equations, models a disease in which a compartment is dedicated to a population that has been exposed to someone infected, but is it not yet infectious to others. Essentially a group of individuals exhibit a delay before indicating symptoms or latent period prior to infection. There are also different transition rates associated with this modification. Lekone and Finkenstadt (2006) developed a stochastic discrete-time SEIR model to estimate parameters such as daily incidence and mortality time series. Their stochastic discrete-time model utilized data from a 1995 outbreak of Ebola in the Democratic Republic of Congo.

Hall et al. (2007) introduced the Susceptible-Exposed-Asymptomatic-Infected-Recovered (SEAIR) model, which includes an asymptomatic population, who are infected but do not show symptoms of the infection. Thus, the transition rates vary because the susceptible population can become exposed then infected or simply asymptomatic. The SIR model has been successfully modified in many applications.

2.3.2 Population Stratification

Epidemic models have stratified populations depending on factors such as age, socioeconomic status, and location. Often, this is in an effort to determine the population that is most likely to spread the disease. Many studies determine that children are the target subpopulation for vaccination to limit the disease spread. For example, Wallinga, Teunis, and Kretzschmar (2006) determined that during an epidemic school aged children and young adults will have the greatest contribution to the future spread of infection due to high incidence rate. A study of NYC influenza morbidity found that there is a greater impact of a specific influenza strain on school-aged children. Generally, this determination could be due to rate of spread, rate of infection, or even social behaviors of a specific subpopulation.

Ajelli, & Litvinova (2017) introduced an age structured SIR model with twelve different age groups. Included in this model are age specific forces of infection that depend on: 1) The number of infectious individuals of a given age, 2) A matrix of contacts regulating the number of contacts with individuals of a given age per unit of time and 3)The transmission rate. This process utilizes a modified SIR framework listed below:

$$\dot{S}_i = -\sum_{j=1}^{n} \beta\, M_{ij} \frac{I_j}{N_j} S_i$$

$$\dot{I}_i = \sum_{j=1}^{n} \beta\, M_{ij} \frac{I_j}{N_j} S_i - \gamma I_i$$

$$\dot{R}_i = -\gamma I_i$$

The equations are governed by these parameters: β is the transmission rate, $\dot{S}_i$ is the number of susceptible individuals in age class i, n is the total number age classes, $\dot{I}_i$ is the number of infected individuals in age class i, $\dot{R}_i$ is the number of recovered individuals in age class i, $\dot{N}_i$ is the number of total individuals in age class i, $M_{ij}$is the matrix of total contacts between individuals of age i and j as derived from the survey, and γ is the recovery rate. Patterns are unique to the region identified by the contact matrix. Regions characterized by contrasting social constructs such as GDP, population structure, and lifestyle will produce different patterns.

Below is a table showing different population stratification groups from several different studies. Each has advantages for its own research purpose.

Table 1 5 different studies with varying quantities and sizes of age groups

Number of Groups	Group Separation	Source
5	0-4, 5-17, 18-49, 50-64, 65+	Worby et al. (2015)
4	1-40, 41-65, 66-80, 81-91	Del Valle, Hyman, Chitnis, (2013)
4	0-4, 5-14, 15-64, 65+	Goeyvarts et al. (2015)
12	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55+	Ajelli and Litvinova, (2017)
8	0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+	Read et al. (2014)

If a community has predefined age groups or data presorted, data analysis should follow that guide and incorporate parameters as able; moreover, the objective of this research is not to add additional burden to a low resource community, but rather work with the available data and integrate established procedures and data with ModSySIR.

2.3.3 Social Contacts

Questionnaires and surveys distributed to different demographics can be a useful tool to determine how to quantify social contacts among individuals and groups (Edmunds, O'Callaghan, Nokes, 1997). A contact is determined by the duration and intensity (proximity) of the interaction. Survey responses can be characterized by factors such as setting (work, social, home, shop, travel, and other) or day (weekday or weekend). Wallinga et al. (2006) used surveys create a social contact matrix M with each element, m_{ij}, that depict the mean number of conversation partners per week in age class i as reported by a participant in age class j. The research took it a step further to estimate age-specific transmission parameters. This matrix N is defined by $n_{ij} = (q\ m_{ij})$, where q is infectivity parameter and N determines transmission rates between age classes. Similarly, Ogunjimi et al. (2009) used a three different methods to estimate a "Who Acquires Infection From

Whom" (WAIFW) matrix based on epidemiological data:$WAIFW\ (i,j) = q \times \frac{m(i,j)}{w_j}$. The matrix is defined by the following parameters: q is the proportionality factor, i & j are age classes, w_j are the number of people in that age class, and $m(i,j)$ is the matrix of average contacts. The rate that participants from class "i" come in contact with a participant in class "j" is defined as $\frac{m(i,j)}{w_j}$. Contact patterns play a role in determining the progression of epidemics. Researchers have analyzed simulated movement of individuals in social roles to determine likelihood of infection (Del Valle, Hyman, Hethcote, and Eubank, 2007). In general, standard compartmental models are useful for theoretical models but don't take into account mechanisms of transmissions and heterogeneities of risk among individuals. An example social contact matrix and its parameters from Del Valle et al. (2007) are listed below:

- Force of infection $\lambda_i$is the rate of disease transmission from infected people in all age groups to suspectibles in age group i.

$$\lambda_i = \sum_{j=1}^{91} \sum_{k=1}^{m} \lambda_{ijk}(t)$$

- $\lambda_{ijk}(t)$ =(Number of contacts per unit time)*(Probability of disease transmission per unit time)*(Fraction of contacts that are infected)
- $\lambda_{ijk}(t) = (\gamma_{ij}(t))(\alpha_i \varepsilon_{jk} P_{ij})(\frac{I_{jk}(t)}{N_j(t)})$;
- $\gamma_{ij}(t)$ number of contacts per unit time
- $\alpha_i$is the susceptibility of a susceptible age group i
- ε_{jk} is the infectivity of an infective stage k of age group j
- $P_{ij}$is the probability of transmission based on the average duration of contacts

> between groups i and j and the fraction of contacts that are infected in group j (2007).

If assuming that a population mixes at random, Keeling and Eames defined the force of infection as the product of transmission rate, effective number of contacts per unit time, and proportion of infectious contacts. The calculation is as follows:

$$\lambda \approx \tau \times \hat{n}\frac{I}{N} = \beta\frac{I}{N}$$

β can be replaced by a matrix of transmission parameters (2005). A contact matrix specific to developing countries will need to be defined for the modified SIR model. Fumanelli, Ajelli, Manfredi, Vespignani, & Merler, (2012) believe that the introduction of human mixing patterns can be used to improve the accuracy of mathematical models. As age change disease susceptibility will also change, so researchers included the additional parameterρ_i, as a coefficient to the $\dot{S}_i$ equation presented by Ajelli and Litvinova (2017) in Section 2.3.2. From H1N1 and seasonal flu data analysis, when ρ_i equal 1, when I is less than or equal to 1.5, otherwise ρ_i equals .5 and R_0 is 1.4. Understanding methods of identify parameters for different models is important when determining the correct parameters for ModSySIR. In fact, Ajelli and Litvinova (2017) found that contact matrices based on surveys allowed them to quantify attack rate specifically by age group. One interesting perspective was presented by Del Valle et al. (2013), in which their model used only 4 major age groups, subsuming all children and young adults into an age group of 1-40; furthermore, the overall population remains unchanged even as changes occur with the different age groups in the infected compartment and is represented mathematically as follows: $N_i = S_i + \sum_{k=1}^{m} I_{ik} + R_i$. Social structure is generally overlooked in epidemic models (Badham and Stocker, 2009). A model addressing these evident gaps needs to be incorporated with SyS data.

2.3.4 Machine Learning in Epidemiology

Machine Learning models are capable of automated pattern recognition as well as evolution and improvement of prediction without additional programming (Shouval et al., 2014). These models lean on consumption of big data classified as training data sets (Kumar, 2018). Machine learning does not replace the need for statistical analysis of data but does provide a means to recognize underlying patterns (Shouval et al., 2014). Key benefits of ML include: automation in changing environments (Quionero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009), efficient pattern recognition for decision support (Shouval et al., 2014). Limitations of using these methods include: barriers to implementation and procedures for accurate parameter identification (Wiens and Shenoy, 2017, Shouval et al., 2014); sample selection bias for both test and training data and inaccurate population proportions (Quionero-Candela, 2009); large data requirements (Pineda et al., 2015); and potential for pattern to generate from random data fluctuations and even overfitting as a result of a limited data set (Shouval et al., 2014).

2.4 Way Forward

This literature review has provided a broad overview of topics affecting SyS research including: modeling techniques, data types and sources, SyS system capabilities, and community characteristics. Gaps in the research and evaluation of SyS systems in low resource areas and subpopulation comparison became evident in this review of literature process. A novel area of research considers both epidemic modeling and the incorporation of SyS data to increase the understanding of subpopulation effects and characteristics. Previous research has utilized SyS observations to estimate parameters of epidemic models (Skvortsov, Ristic, and Woodruff, 2010). A comprehensive study should be dedicated to study how variations of data based on population characteristics can affect the ability to detect an outbreak. Low resources countries need "practical tools", such as a model to understand potential impact, feasible mitigations, and planning guidelines for pandemics (Oshintani, Kamigaki, and Suzuki, 2008). The very nature of SyS allows for an earlier warning compared to laboratory confirmed data. Considering research gaps, and potential opportunities, the objectives of this research is the following:

- Identify which subpopulation best represents the encounters of the SyS data set.
- Determine subpopulation with leading indicators of an impending outbreak and which subpopulations would result in missing an increase in overall ILI activity.
- Implement a mathematical model demonstrating community behavior among subpopulations along with SyS ILI data.
- Determine how different subpopulations compare with the ability to accurately predict the peak of an ILI outbreak.

The previously referenced modified SIR models added compartments and additional parameters to make the theoretical model more realistic and representative of the complex nature of epidemic or disease behavior. One challenge of implementing a subpopulation based epidemic model is incorporating a "realistic contact structure" (Ball et al. 2015). Considering the rate in which individuals or groups interact with one another is an additional opportunity to modify SIR models to further depict an accurate representation of social behavior. In the case of ILI, our syndrome of interest, individuals can experience an asymptomatic period and variation in the severity of their illness. Quantifying parameters of transmission that fit various communities requires preparation and community engagement or outreach prior to the implementation of public health outbreak guidance. Engagement such as a social behavior study or survey inquiring about contacts such as age, location, frequency, or proximity enables practical transmission parameters. Public health officials can intercept an outbreak efficiently by both identifying high risk populations and studying their contact patterns and pinpointing necessary behavioral changes (Del Valle et al., 2013). In the initial stages of an outbreak social behavior will remain consistent, as opposed to conscious social distancing. Ajelli and Litvinova (2017) reviewed both heterogeneous and homogenous mixing models while one requires transmission rate and the later a contact matrix. For our purposes, we will assume homogeneous mixing and incorporate contact pattern behavior into our parameter for our transmission rate. Consider Goeyvaerts et al. (2015) presented a social contact hypothesis such that “Age-specific transmission rates are directly proportional to age specific rate of making social contact”. The research proposes that a community can reuse a contact matrix from a community with similar characteristics (size, economic status, health infrastructure, or population make up). Low resource or developing communities should focus data collection effort on a subset of the population for the following reasons:

- Early indication from an identified subpopulation
- Transmission parameters account for social contacts

- Epidemic model consuming data has realistic characteristics

The research will use a representative SIR model with transmission parameters to identify early indicating subpopulations. Chapter 5 discusses implementation of the approach and application of ModSySIR. Finally, the model will incorporate a portion of population represented by SyS data. From this evolved model, community specific public health guidelines and decisions, such as plans for outbreak prevention, intervention, and response, can be drafted. Chapter 3 will further detail how historical SyS data and the SIR model will be used to draw conclusions to improve limited resource public health surveillance.

Chapter 3: Research Approach

3.1 Overview

This effort will focus on subpopulation characterization of NYC SyS data and a model to identify how these subpopulations provide different predictions. The characterization involves data analytics and summary statistics for each flu season. To best evaluate the SIR model, real SyS data will be utilized. The data used is publicly available from the New York City (NYC) Health Department. Hopkins (2012) stated SyS systems can provide insight into the geographic distribution of a syndrome, subpopulation groups most affected, syndrome behavior, and hospital capacity impact. Thus, this analysis will segregate the NYC data into age and geographic subgroups and then assess the data for general trends and anomalies. MATLAB will be used to evaluate the system of equations for ModSySIR and compare to actual data of confirmed diagnostics and trends. The following are the main steps of the proposed research approach:

Step 1: Collect SyS Data

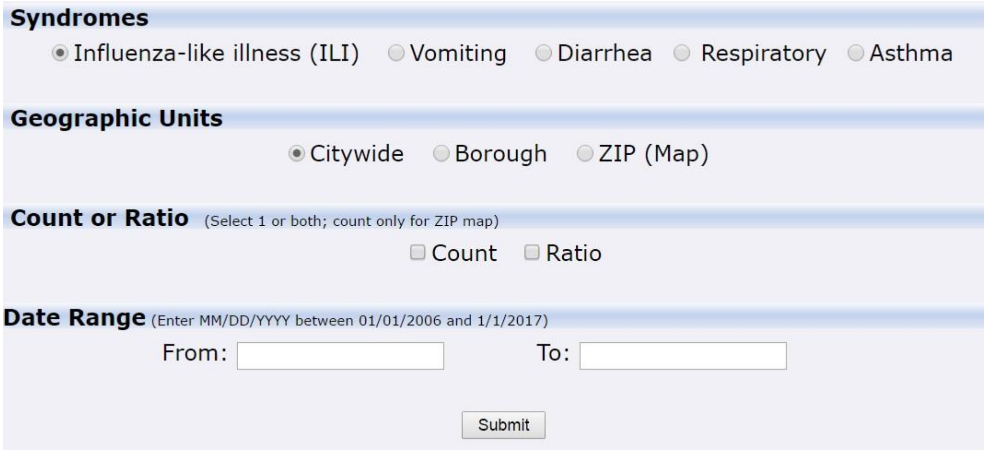
Step 2: Analyze Data

Step 3: Adapt SIR Model for Transmission Among Multiple Subpopulations

Step 4: Analyze Model Results

3.2 Collect SyS Data

As stated in the literature review, there are many data types collected for the varying Syndromic Surveillance Systems used around the world. The SyS data analyzed for this effort was collected from New York City's (NYC) Health Department. NYC collects patient encounter data from all fifty-three emergency departments (EDs) in the city. NYC EDs send syndromic data to the health department daily. The system counts individual patient encounters; therefore, patients that have multiple visits will be counted multiple times in the data set. The interactive data organizer, Epiquery, provides data for five different syndromes: asthma, diarrhea, influenza-like- illness (ILI), respiratory, and vomiting. This online application can be found at https://a816-healthpsi.nyc.gov/epiquery/Syndromic/. This data can be exported to Microsoft Excel for further analysis.



Syndromes

● Influenza-like illness (ILI) ○ Vomiting ○ Diarrhea ○ Respiratory ○ Asthma

Geographic Units

● Citywide ○ Borough ○ ZIP (Map)

Count or Ratio (Select 1 or both; count only for ZIP map)

□ Count □ Ratio

Date Range (Enter MM/DD/YYYY between 01/01/2006 and 1/1/2017)

From: ______ To: ______

Submit

Figure 2 The user's view of necessary inputs for SyS data from Epiquery (New York City Department of Health and Mental Hygiene, 2015)

Epiquery data is structured by geography, age, and syndrome and is available from 2006 to present day. For this effort, the data set was collected in 2016 and analyzed for ILI encounters over a period of eight years: January 1, 2008, to December 31, 2015. The available data currently may vary as additional data is ingested. Statistically, flu trends are characterized by season, not the calendar year. Thus, the flu season trend (mid-October to the end of April) were the basis for summary statistics as opposed to the entire calendar year (Fleming et al., 2000; Cooper et al., 2007). Thus, the data describes seven flu seasons (from 2008-2009 to 2014-2015). Geographically, the data was analyzed at the citywide and at the borough level. According to Epiquery, the borough is assigned by the residential ZIP code of the patient. If the ZIP code is missing from patient data, a likely ZIP code of residence is assigned based on the ED location and demographic characteristics. The age groups are predetermined and organized by Epiquery prior to analysis: 0-4 years, 5-17 years, 18-64 years, and 65 years and older. If the age is missing from patient data, the patients are omitted from the age-specific categories, but included in the "all ages" group data. The data set of interest has over 600,000 patient encounters over eight years (1/1/2008 to 12/31/2015).

The data does have some limitations applicable to this data analysis. The Staten Island produced much less encounter data than other boroughs. This is simply due to lack of emergency department chief complaint reporting prior to 2016 in that borough. NYC SyS system currently captures 100% of all ED visits in the city; however, not all resident choose to seek care in EDs, so the entire population is not covered. For all boroughs, data is unavailable for three dates: 6/13/2014, 9/5/2014, and 12/25/2014. These deficiencies are acknowledged by the NYC health department. It should be noted that this could have potential implications on a model utilizing this data set because one age group may be prone to revisits which in turn could skew the population sample size. Despite these limitations, the data set is robust and flexible for analysis.

Clinical data is collected at both the regional and national level in a joint effort led by the CDC. More information about clinical data used for this study can be found at https://www.cdc.gov/flu/weekly/overview.htm#Viral and the interactive data set can be found at https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html (CDC, 2018). Many other research efforts have utilized this nationwide data source for ILI related research (Santillana et al., 2015) Similar to the NYC SyS, this data also spans from

2008-2016. The peaks found in this data set will be considered to be a baseline for further data analysis.

3.3 Analyze Data

To generate summary statistics, trends, and population characteristics, the daily patient encounter data was aggregated to form weekly values, and the following statistics are calculated:

- The fraction of each age group in the overall population
- The fraction of each borough in the overall population
- Flu season encounters for each age group (All Years: 0-4, 5-17, 18-64, 65+)
- Flu season encounters for each borough (All Years: Bronx, Brooklyn, Manhattan, Queens, Staten Island)

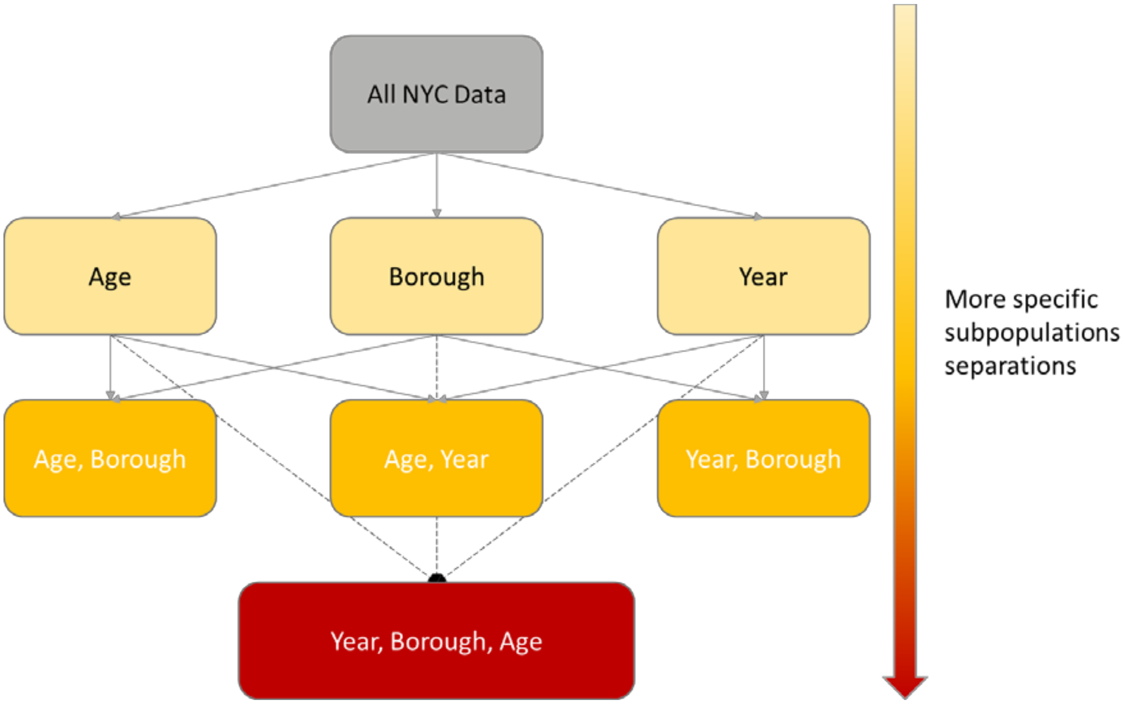


Figure 3 Various methods to organize the data for analysis

For each flu season, the data was rescaling or normalized by finding the average weekly number of reported encounters and dividing each week's number of reports by that year's average. This process was to ensure that by normalizing the all of the data, the behavior of subpopulations with less encounters would still be captured visually and maintain statistically relevant. Then the data was analyzed by year (flu season), age group, and borough. Leading indicators were determined by looking at five week periods leading to a peak in encounters and identifying normalized increases in reports. This identified an age group and/or borough that is more susceptible to early infection. The correlation of a subgroup encounters and the total population encounters was calculated considering a one and two week lag in data.

The following section steps through the summarized data analysis process. The key steps include the organization, normalization, correlation, and characterization of subpopulation data. First the data is organized into subsets by age, location, and flu season. The data is aggregated every seven days to determine descriptive and summary statistics for each data set. Next, the data is normalized by determining average values for each seven day period. Summary statistics are recalculated. Supporting charts can be found in Section 4.1 as well as the appendix. Then Pearson's Correlation Coefficient and Root Mean Squared Error is calculated to identify SyS ILI subpopulation activity leading to a peak in the data set and compared to the total population SyS ILI activity. The same function is also used to determine subpopulation correlation considering the total population data set but subpopulations by both lagging 1 week and 2 weeks, respectively. Thus, the data from weeks preceding a peak would be shifted in this comparison to see if a subpopulation show a leading indication of an upcoming peak in the total population. Based on the correlation values, the groups with the highest correlation values are identified. The data analysis comprises the following key steps:

1. Organize Data
    a. Yearly data and Flu season
2. Identify and compare peaks in National Clinical Data and NYC Syndromic Surveillance Data
    a. Include Preceding 7 weeks
3. Utilize National Clinical Data as benchmark for data likeness assessment: Root Mean Squared Error (RMSE) and Pearson's correlation

Section 4.2 displays how this process translates to identification of the subpopulation with highest correlation.

3.4 Understanding the Metrics

A variety of metrics, including root mean squared error (RMSE) are needed to determine model performance (Chani and Draxler, 2014). RMSE is the square root of the variance of the residuals; furthermore, a lower RMSE value indicates that observed and predicted values are closer to one another. In this case, the observed values consist of the SyS data and the predicted values are the nationwide clinical flu data. Differing from the correlation coefficient, RMSE is a measure of fit for a model or data sets.

Pearson's correlation coefficient measures the strength of linear correlation between data sets. In this case, using Pearson's to measure correlation for an entire flu season is not the best measure. Rather, the subset of the data such as time periods leading to a peak lend themselves to this type of evaluation. Researchers used Pearson's Correlation calculation to evaluate weekly data from three sources over five flu seasons: Google Flu Trends, rates of ILI, and surveillance for laboratory-confirmed influenza

(Ortiz et al., 2010). Researchers compared influenza data mined from Twitter to data from the Infection Disease Surveillance Center in Japan to determine correlation (Aramaki, Maskawa, and Morita, 2011). The performance of Google Flu Trends data, which are weekly estimates based on internet search terms, was compared to U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) data. Specifically, the two measures this study utilized in data analysis, correlation and RMSE, was determined by comparing Google Flu Trends model estimates and ILINet data within four time periods ranging from Mar 2009–Dec 2009 (Cook, Conrad, Fowlkes, and Mohebbi, 2011). One goal of that study was to evaluate how internet search behavior varied depending on the time period. There are further uses of these calculation in Machine Learning applications. Santillana et al. (2015) compared three machine learning algorithms which combined ILI estimates from various sources and the performance of the predictors was assessed using RMSE Pearson’s correlation calculation as well as other assessments. Similarly, the different models were assessed as researchers evaluated the forecasting performance using RMSE (Xu, Gel, Ramirez, Nezafati, Zhang, 2017; Brownstein et al., 2017).

The Centers for Disease Control and Prevention (2010) evaluated the impact of H1N1 on different races and ethnicities. This is an example of the research that considered how influenza effects subpopulations based on a weekly data collection. Similarly, this case study looks at subpopulations, both age and location, and metrics to compare behavior of each subpopulation.

3.5 Adapting SIR Model for Transmission Among Multiple Subpopulations

The ModSySIR will account for transmission by incorporating social behavior with SIR and implement a proportion of the SyS data. SIR system of equations and additional transmission parameters and other potential modifications will be evaluated with SyS data in comparison to real data. ModSySIR model defines a parameter for transmission among subpopulations.

As previously stated in the literature review, as individuals transition between compartments S, I, and R, and the overall population (N) remains unchanged. Keep in mind, that mortality and birth are neglected in this approach; however, there are more complex extensions of the SIR model that account for these factors. Let S, I, and R be the number of susceptible, infected, and recovered individuals; therefore, N=S+I+R. Traditionally initial conditions are unique to the population, but are notionally the following: when N=1, S(0)=.99, I(0)=.01, and R(0)=0 for the discrete SIR model. N will always represent the total population, whether the value is 1, 100, or the actual population. If N is the actual population, then accommodate for that change by dividing β by N, analogous to discussion in section 2.3.1. The transmission parameter or social contact matrix is predetermined considering options for literature, community behavior, and SyS data.

As previously mentioned in section 2.3.2 the model presented by Ajelli & Litvinova (2017) considers: 1) The number of infectious individuals of a given age, 2) A matrix of contacts regulating the number of contacts with individuals of a given age per unit of time and 3) The transmission rate.

This process utilizes a modified SIR framework:

$$\dot{S}_i = -\sum_{j=1}^{n} \beta\, M_{ij} \frac{I_j}{N_j} S_i$$

$$\dot{I}_i = \sum_{j=1}^{n} \beta\, M_{ij} \frac{I_j}{N_j} S_i - \gamma I_i$$

$$\dot{R}_i = -\gamma I_i$$

Details on the notation and parameters can be found in section 2.3.2. This framework can also be found in a paper by Ajelli & Litvinova (2017). Social contacts are determined by location, duration, and intensity. For our purposes, these have been predetermined in prior research.

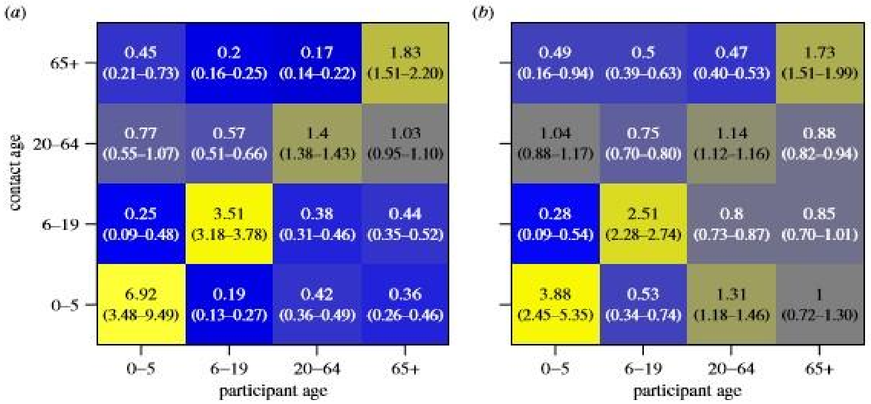


Figure 4 Two age-based mixing matrices for (a) total number or contacts and (b) total contact duration. 95% Confidence intervals are indicated in parentheses. (Read et al., 2014)

An example of a quantified contact matrix is in Figure 4. The yellow colors indicate a greater level of mixing between age groups while bluer colors show less mixing. Data was collected via self-report and responses were statistically analyzed to create the contact matrices. Based on behavior of individuals and mixing patterns, these contact matrices can be simplified to transmission parameters when appropriate. Fumanelli et al. (2012) found that synthetic contact matrices helped to improve the accuracy of mathematical model prediction which aligns with the goals of the research.

Del Valle et al. (2013) explored a multigroup SIR model with considerations of age dependences and various mixing patterns. We're assuming homogeneous mixing so our matrix can be substituted for a transmission parameter. We will identify that transmission parameter based on reproduction number found in Ajelli and Litvinova's (2017) work.

Table 2 lists the parameters and variations used for the ModSySIR model:

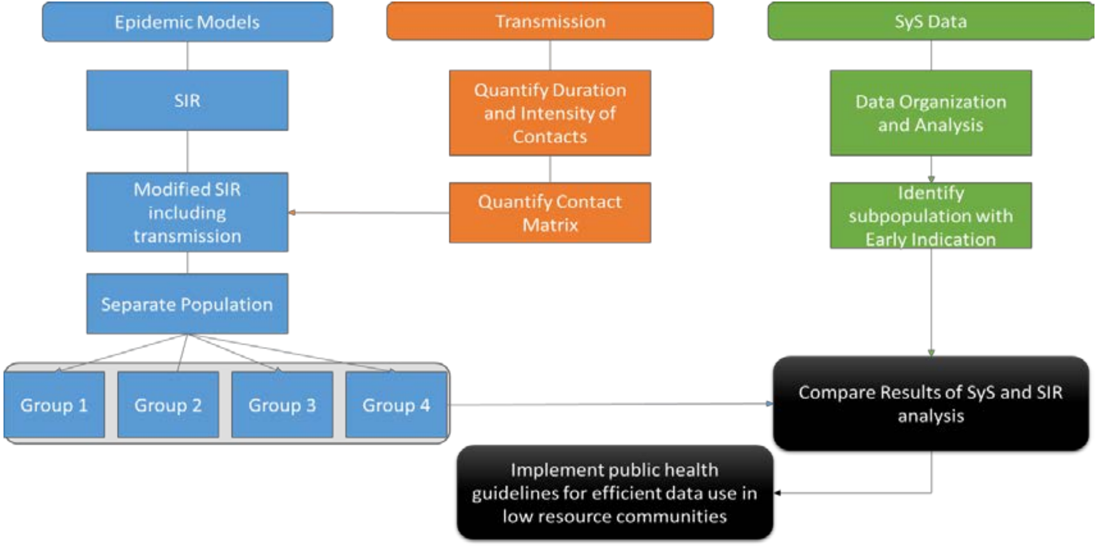
Table 2 Parameters in use for the ModSySIR model and the associated sources or derivation.

Parameter	**Value**	**Source**
R_o infection rate/recovery	1.3 – Seasonal 1.6 – Pandemic	Ajelli & Litvinova
Transmission Coefficient	ρ_i = .5	Ajelli & Litvinova
Subpopulation	Refer to Section 4.3	NYC Data & Lab Data
Infection Rate	Refer to Section 4.3	NYC Data & Lab Data
Recovery Rate	Extrapolate from R_o	

Infection rate is determined from SyS data: combined encounters leading to the peaks (5 preceding weeks) is divided by the yearly encounters for the total population. Finally, these outputs will be compared to available SyS data for each subpopulation, SyS data considering the entire population, and confirmed clinical data. This is to determine model utility as well as the subpopulation with the best fit to the SIR model.

Overall, this customizable SIR model incorporates community data and helps to identify SyS subpopulation with the ability to indicate impending disease behavior. This process is substantiated in MATLAB and the group that provides displays the earliest indication of an ILI outbreak through decrease in susceptible population and increase in infected population is identified. These results are compared to the completed data analysis for each flu season. Figure 5 gives an overview of how the methodology is employed. Chapter 4 gives a logical explanation of this process and application of approach on a specific subset of data. Chapter 5 presents the results of this research approach.

Figure 5 A high level illustration of the research method and intermediary steps


Epidemic Models
Transmission
SyS Data
SIR
Quantify Duration and Intensity of Contacts
Data Organization and Analysis
Modified SIR including transmission
Quantify Contact Matrix
Identify subpopulation with Early Indication
Separate Population
Group 1
Group 2
Group 3
Group 4
Compare Results of SyS and SIR analysis
Implement public health guidelines for efficient data use in low resource communities

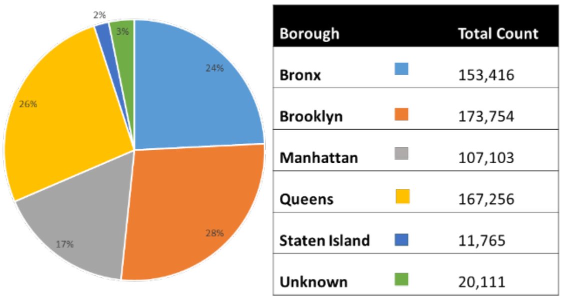
Chapter 4: Results

This chapter presents the results: Section 4.1 summarizes the SyS data from New York City; Section 4.2 provides a detailed looked at the data analysis process and resulting insights specific to the 2014-15 flu season; these results include identification of leading indicators and correlations of subpopulations; This section also summarizes data analysis of the entire SyS data set and flu lab data and highlights what data will be pulled into ModSySIR. Complete results for all years of the data analysis can be found in the appendix. Section 4.3 presents the modified SIR model and associated parameters based on previous results. In Section 4.4, applicable results are presented for implementation of the ModSySIR model and translation of ModSySIR results.

4.1 Summary of SyS Dataset

The following graphs and tables give a quick overview of the NYC SyS ILI data considering all encounters and years of data. The initial graphic depicts multiples ways that the data is separated and analyzed. The date range of data collection is 1/1/2008 to 12/31/2015. The data is evenly distributed between boroughs with the exception of Staten Island, which produces fewer reports of encounters. The age distribution shows that that there is little representation of the 65+ age group. There is an influx of ILI encounters in 2009 due to the severity of H1N1, while all other years fluctuate between 65,000 – 80,000 reports.

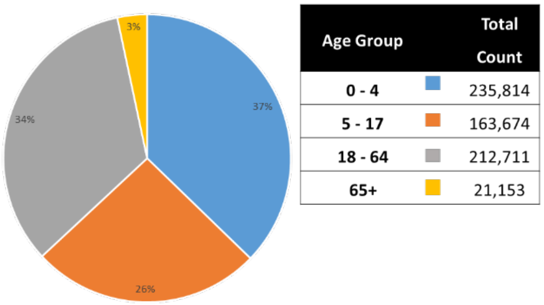
Borough Distribution



Borough		Total Count
Bronx	🟦	153,416
Brooklyn	🟧	173,754
Manhattan	▨	107,103
Queens	🟨	167,256
Staten Island	🟦	11,765
Unknown	🟩	20,111

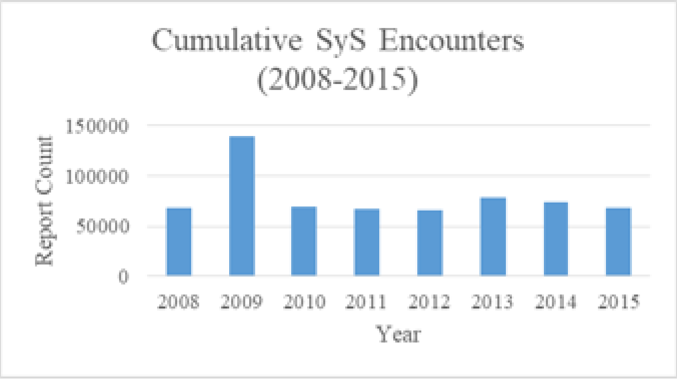
Figure 6 From the SyS data, the pie chart represents the distribution of NYC's population separated by borough. The five boroughs are Bronx, Brooklyn, Manhattan, Queens, and Staten Island. The unknown category is for encounters that did not have a zip code when recorded.

Age Distribution



Age Group		Total Count
0 - 4	🟦	235,814
5 - 17	🟧	163,674
18 - 64	▨	212,711
65+	🟨	21,153

Figure 7 From the SyS data, the pie chart represents the distribution of NYC's population separated by age. The four age groups are 0-4, 5-17, 18-64, and 65+.



Year	Total Count
2008	68,998
2009	139,321
2010	69,557
2011	67,355
2012	66,235
2013	79,245
2014	74,464
2015	68,230
Total	633,405

Figure 8 From the SyS data, the chart presents yearly cumulative encounters from 2008 to 2015. This is inclusive of the entire data.

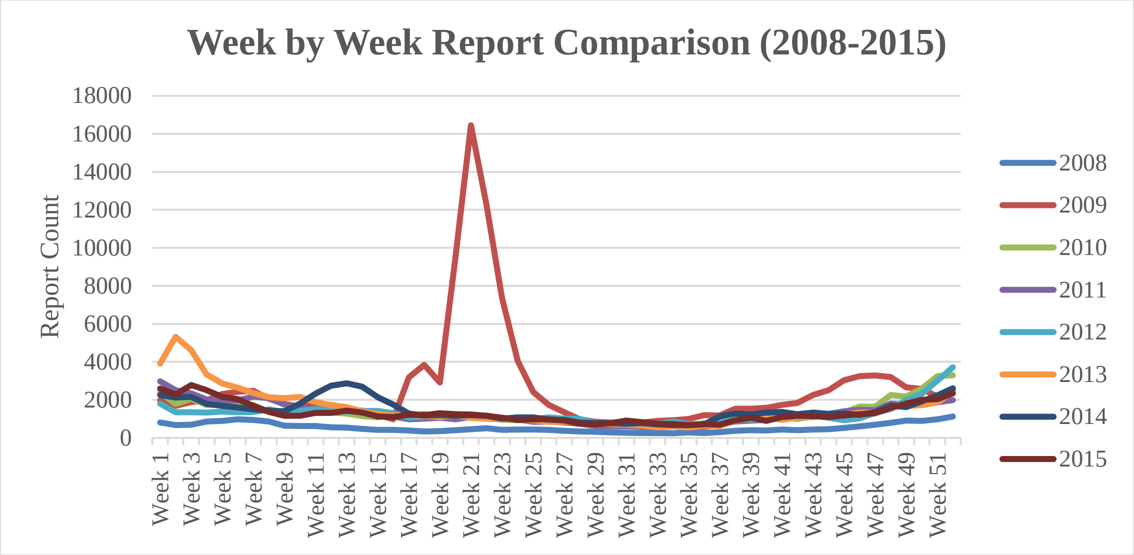


Figure 9 A comparison of ILI encounters for week 1-53 reports for each year. 2009 showed a unique increase of encounters outside of the traditional flu season.

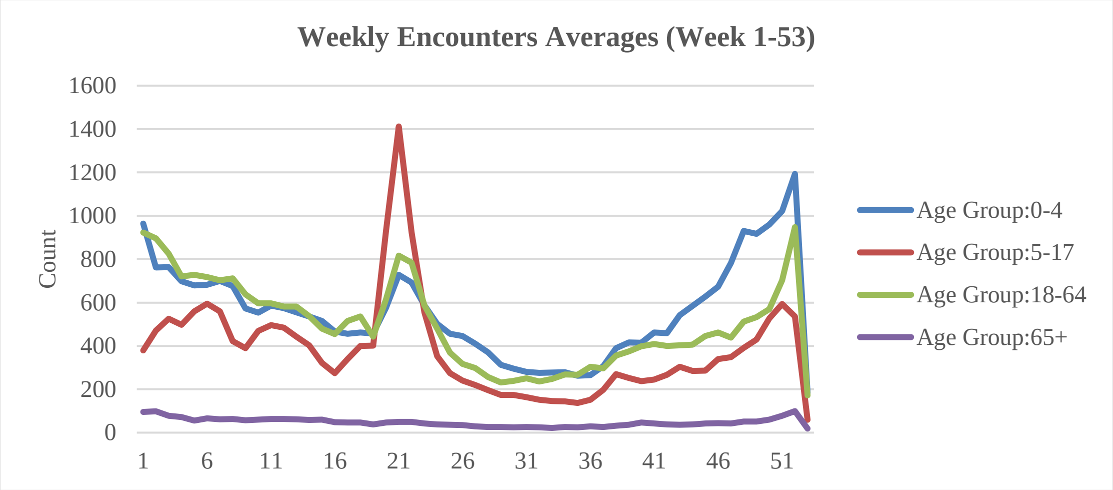


Figure 10 A comparison of activity of each age group for week's 1-53 reports for 2008-2016

Figure 9 shows how weekly averages can be largely skewed by the H1N1 outbreak in which there was a great increase in encounter reporting around week 21. Each subpopulation is also analyzed in depth. The data is separated by flu season rather than the normal calendar year. Each flu season will have absolute and normalized encounters for both age groups and boroughs. The 2014-2015 flu season SyS encounter and lab data is examined in section 4.2.

4.2 Data Sample Summary

This section scrutinizes the results of the 2014-2015 flu season only, rather than a full calendar year. Summary statistics, correlation, and RMSE of SyS data, and Clinical Lab influenza data for various time periods is explored. The implications of these results will be discussed as the data is presented.

4.2.1 SyS Data

As previously explained in Chapter 3, there are over 600,000 encounters to review from the NYC SyS Data Set. The normalized data from the 2014-15 flu season organized by both age and location are show side by side in table 3. The color coding can be used as a heat map to visually recognize the variation in the encounters.

Table 3 2014-2015 Sys Data separated into 4 age groups and 5 boroughs. If residential data is unavailable, that encounter is categorized as unknown.

	Encounters for Age Groups				**Encounters for Borough**					
Week	0 - 4	5 - 17	18 - 64	65+	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Sum
20	448	300	367	53	329	254	194	341	21	1168
21	484	294	366	59	294	331	185	333	18	1204
22	457	247	285	39	258	244	168	320	15	1028
23	434	263	315	37	245	293	165	303	16	1049
24	484	292	303	46	285	300	186	316	8	1126
25	434	242	253	27	262	222	125	311	11	957
26	430	209	261	33	224	229	144	301	8	933
27	375	197	222	32	180	210	140	254	14	826
28	304	187	235	31	193	187	118	225	5	757
29	287	177	259	44	204	185	139	201	18	767
30	295	192	232	32	194	199	140	180	17	751
31	325	158	300	26	197	164	158	249	14	809
32	265	146	231	35	169	176	98	198	12	677
33	273	121	276	31	171	180	112	200	12	701
34	256	136	241	25	159	158	115	181	16	658
35	249	168	268	32	192	155	134	203	13	717
36	399	320	268	29	249	253	183	291	13	1016
37	519	393	397	33	319	402	221	350	25	1342

38	482	313	396	33	279	330	211	343	28	1224
39	495	282	449	59	323	320	210	377	24	1285
40	564	295	454	55	286	370	264	379	24	1368
41	458	340	451	43	291	343	242	345	25	1292
42	544	347	399	37	324	338	243	356	27	1327
43	543	308	362	47	296	371	215	322	21	1260
44	606	301	307	40	272	374	215	342	17	1255
45	543	298	344	38	301	327	238	317	17	1223
46	663	307	334	41	346	335	241	360	18	1345
47	837	344	400	40	371	404	267	513	18	1621
48	788	358	435	48	389	482	249	438	26	1629
49	828	475	487	81	464	460	340	523	32	1871
50	883	523	598	84	525	503	336	612	32	2088
51	1118	536	850	134	639	622	429	768	62	2638
52	899	423	1025	159	638	607	433	693	62	2506
1	674	593	943	156	568	632	413	627	58	2366
2	820	855	983	152	710	644	481	845	56	**2811**
3	721	713	888	165	632	664	414	641	66	2488
4	598	561	810	134	559	561	298	556	51	2103
5	617	535	775	116	569	480	376	531	44	2043
6	528	308	646	102	380	411	255	460	34	1584
7	423	239	572	89	325	334	238	348	35	1323
8	363	243	453	68	302	293	184	296	19	1127
9	431	286	452	57	315	288	221	339	25	1226
10	478	336	438	60	330	321	190	407	24	1312
11	440	378	430	64	322	317	182	447	13	1313
12	545	388	461	64	338	334	214	513	27	1458
13	476	281	452	70	310	306	172	439	20	1280
14	436	211	426	56	279	250	158	392	20	1129
15	418	270	347	46	268	269	164	338	16	1081
16	466	305	389	48	287	289	191	389	19	1208
17	477	353	363	56	286	262	178	453	25	1249
18	482	384	325	52	265	299	206	427	18	1243
19	492	348	361	53	320	324	155	408	12	1254

Basic characterization of the raw data is summarized and charted below. In general, children 0-4 and the borough of Queens represent most encounters and have the highest weekly average. Age 65+ and the borough of Staten Island has the least number of encounters.

	Average	Range	Max	Min
0-4	516	869	1118	249
5-17	328	734	855	121
18-64	440	803	1025	222
65+	61	140	165	25
All	1346	2153	2811	658

	Average	Range	Max	Min
Bronx	331	551	710	159
Brooklyn	338	509	664	155
Manhattan	222	383	481	98
Queens	390	665	845	180
Staten Island	24	61	66	5
Unknown	39	98	118	20
Boroughs	1346	2153	2811	658

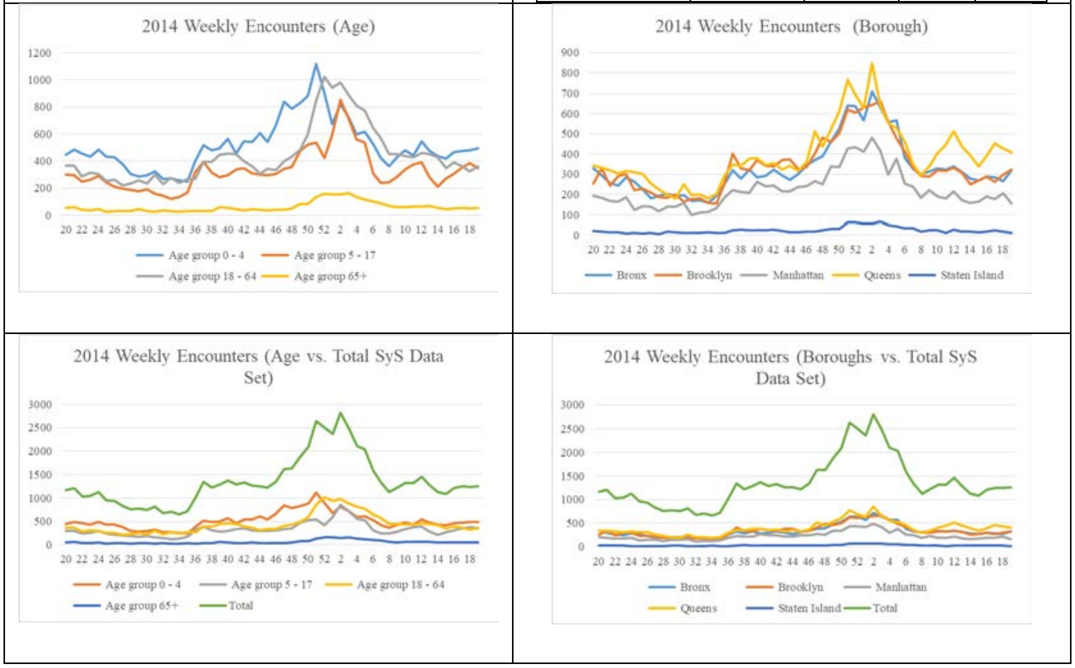
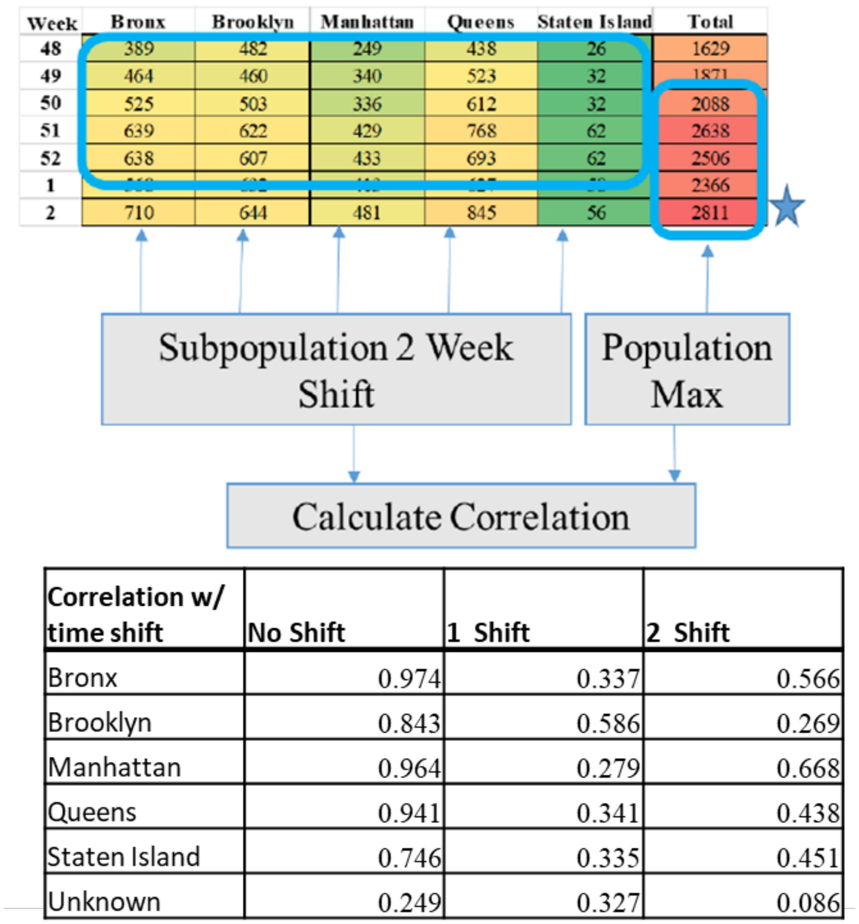


Figure 11 Summary Tables and Charts of 2014-15 Flu Season SyS data for subpopulations

The peak week for the 2014-2015 Flu Season based on SyS data is week 2 and the value is 2811 encounters. All subpopulations generally peak around the same time; the supporting charts shown in figure 11 illustrate encounters for various subpopulations across flu season show an increase in encounters for most subpopulations as they approach week 2. Once the peak of the SyS Data is identified, each subpopulation is examined to see which is the best anticipates the impending increase in encounters. The goal is to identify the subpopulation with the maximum correlation coefficient at considering data analysis at 3 different time period shifts: No shift, 1 week lag, and 2 week lag.

The summary data and correlation data charts are summaries of correlation analysis from 3 different time periods. The goal of this step is to characterize SyS data by identifying which boroughs and age groups have encounter behavior most similar to the overall population. Figure 12 shows how the data is shifted for each time period which leads to the correlation. Pearson's Correlation Coefficient allows the comparison to compare five weeks of data within each subpopulation to the preceding 5 weeks leading to the peaks in total population. These comparisons occurred for each subpopulation 1 and 2 weeks prior to total population ILI peak. Considering all subpopulations of both boroughs and age groups, the subpopulation with the maximum correlation coefficient at each time shift is identified as well as those with r greater than .8. Once the peak is identified and the subpopulation correlations are calculated, the values are charted for a visual comparison. The correlation values for all years, ages, and boroughs are summarized in the appendix. Figure 13 illustrates how the correlation and RMSE of subpopulations is identified in various scenarios for both age and location groups.



Week	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Total
48	389	482	249	438	26	1629
49	464	460	340	523	32	1871
50	525	503	336	612	32	2088
51	639	622	429	768	62	2638
52	638	607	433	693	62	2506
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2366
2	710	644	481	845	56	2811

Correlation w/ time shift	No Shift	1 Shift	2 Shift
Bronx	0.974	0.337	0.566
Brooklyn	0.843	0.586	0.269
Manhattan	0.964	0.279	0.668
Queens	0.941	0.341	0.438
Staten Island	0.746	0.335	0.451
Unknown	0.249	0.327	0.086

Figure 12 An example of correlation coefficient of NYC boroughs for the 2014-2015 flu season. The peak is starred at week 2 and the 5 preceding weeks are included. The boroughs with the highest correlations for no shift, 1 week shift, and a 2 week shift are: Bronx, Manhattan, Queens, and Brooklyn, respectively.

For this example, the Bronx has the best matches the encounter behavior of the total SyS dataset; however, Brooklyn, Manhattan, and Queens are also significant with correlation values where r > 0.8. This process is completed for every flu season. The short summary of correlation and RMSE results are below. The rest of the results can be found in the Appendix. Using RMSE, the Bronx still has the best match to the total SyS dataset; however, Brooklyn, Manhattan, and Queens are also significant with correlation values where Correlation Coefficient r > 0.8. The RMSE calculation is also completed for every flu season.

	Age	Borough

Correlation

Maximum Correlation Values (Age)

	No Shift	1 Week	2 Week
Year	0.941	0.875	0.872
Flu Season	0.918	0.768	0.832
Peak	0.729	0.762	0.666

Maximum Correlation Values (Borough)

	No Shift	1 Week	2 Week
Year	0.988	0.923	0.854
Flu Season	0.987	0.878	0.752
Peak	0.974	0.586	0.668

Summary of Age Groups with Correlation above .8

	No Shift	1 Week	2 Week
Year	All - Max 18-64	All - Max 0-4	0-4
Flu Season	18-64, 65+, 5-17	★ N/a	0-4
Peak	N/a	N/a	N/a

Summary of Boroughs with Correlation above .8

	No Shift	1 Week	2 Weeks
Year	All - Bronx Max	All - Brooklyn Max	Manhattan, Brooklyn, Bronx, Queens
Flu Season	All - Bronx Max	Queens, Brooklyn, Bronx, Manhattan	N/A
Peak	★ Bronx, Manhattan, Queens, Brooklyn	N/A	N/A

RMSE

RMSE	0-4	5-17	18-64	65+
Flu Season	0.181	0.164	0.166	0.307
Peak - 2 Week	0.296	0.491	0.526	0.579
Peak - 1 Week	0.811	1.205	1.066	1.224
Peak – No Shift	0.328	0.376	0.292	0.526
Lowest RMSE	**Group**	**RMSE Value**		
Flu Season	5-17 ★	0.164		
Peak - 2 Week	0-4	0.296		
Peak - 1 Week	0-4	0.811		
Peak – No Shift	18-64	0.292		

RMSE	Bronx	Brooklyn	Manhattan	Queens	Staten Island
Flu Season	0.062	0.085	0.097	0.099	0.290
Peak - 2 Week	0.343	0.350	0.328	0.399	0.596
Peak - 1 Week	0.990	0.882	0.998	1.008	1.168
Peak – No Shift	0.047	0.119	0.073	0.085	0.504
Lowest RMSE	**Group**	**RMSE Value**			
Flu Season	Bronx	0.062			
Peak - 2 Week	Manhattan	0.328			
Peak - 1 Week	Brooklyn	0.882			
Peak – No Shift	Bronx ★	0.047			

Figure 13 Summary of Correlation and RMSE Calculations for both and age and borough for the 2014-2015. Heat maps indicate the range of values for the calculations. Stars shows similarities between Correlation and RMSE results.

The first row of figure 13 presents maximum correlation values for various time periods and different time shifts. The second row is a qualitative summary of all subpopulations with correlation measures above .8 for those time periods and shifts. The flu season correlation data is a more relevant concern for this research as opposed to an entire year of data. Age groups 5-17, 18-64, and 65+ show that these subpopulations are good indicators in real time of flu season encounters for the entire SyS data set; however, if the subpopulation data shifts 1 week earlier, there is no age group subpopulation showing anticipation of an impending outbreak. Similarly when analyzing the two week shift, only one subpopulation 0-4, shows a leading indicator. The peak related calculations are key to identifying ModSySIR parameters; unfortunately the data correlation summary for age group shows no subpopulation in sync or shifted with a predictive indicator of the upcoming peak. On the contrary, 4 boroughs show good correlation for peaks at that time shift. The third row identifies RMSE for each subpopulation at the subpopulation that has the lowest RMSE for the entire flu season. Finally, the last row identifies RMSE for each age group focusing on the peak and the preceding 5 weeks at various time shifts found in this particular flu season. Even though there is no clear subpopulation with outbreak indication, there is a lot to learn from this analysis. Moving forward, other analytics will be performed to interpret the data. The blue stars found on this chart indicate where both correlation and RMSE calculations provided similar results. To better understand what these results mean, this process should be explored across the entire data set to identify trends in subpopulations.

4.2.2 Comparison of Confirmed Lab Data and SyS Data

Flu View ILI Net Nationwide (Centers for Disease Control and Prevention 2018) data was used and a summary graphic and a table of statistics can be found in Figure 14 and Table 4. Similar to SyS Data, there is an increase in the 2009-10 flu season due to the H1N1 flu strain. The peak week identified for each flu season will now be the baseline moving forward for all comparisons of SyS and Clinical data including the formulation of model parameters.

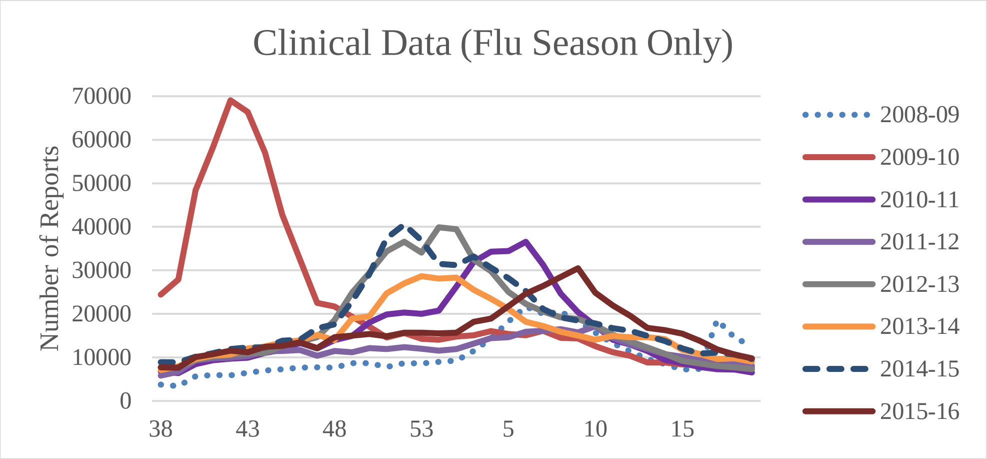


Figure 14 Influenza Clinical Lab Data for 2008-2016

Table 4 Summary of Influenza Clinical Lab Data for each year 2008-2016 including weekly averages and peak values and the week at which they occur

	2008-09	**2009-10**	**2010-11**	**2011-12**	**2012-13**	**2013-14**	**2014-15**	**2015-16**
Year Total	421398	**1007169**	661283	493401	753776	654609	789209	668119
Weekly Average	7951	**19003**	12477	9309	14222	12351	14891	12606
Peak Value	21667	69068	36567	17150	39896	28654	40533	30481
Peak Week	6	42	5	9	52	52	**52**	8

Table 4 provides a summary of clinical nationwide data which include yearly totals, weekly averages, peak values, and the week at which the peak occurs. The 2014-

15 flu season will be investigated to illustrate the data analysis and parameter identification process. The peak week in 2014-15 flu season is week 52 and correlation and RMSE calculations will revolve around that week and the time period including the increase the up to flu season peak. Figure 15 compares age and borough encounters in 2014-15 flu season to visualize behavior of each subpopulation. For the SyS data the peak occurs at week 2. Similarly, the likeness of each age group and borough to clinical data, shifted populations back 1 & 2 weeks, is evaluated. This visualization led to correlation and calculation in figure 16.

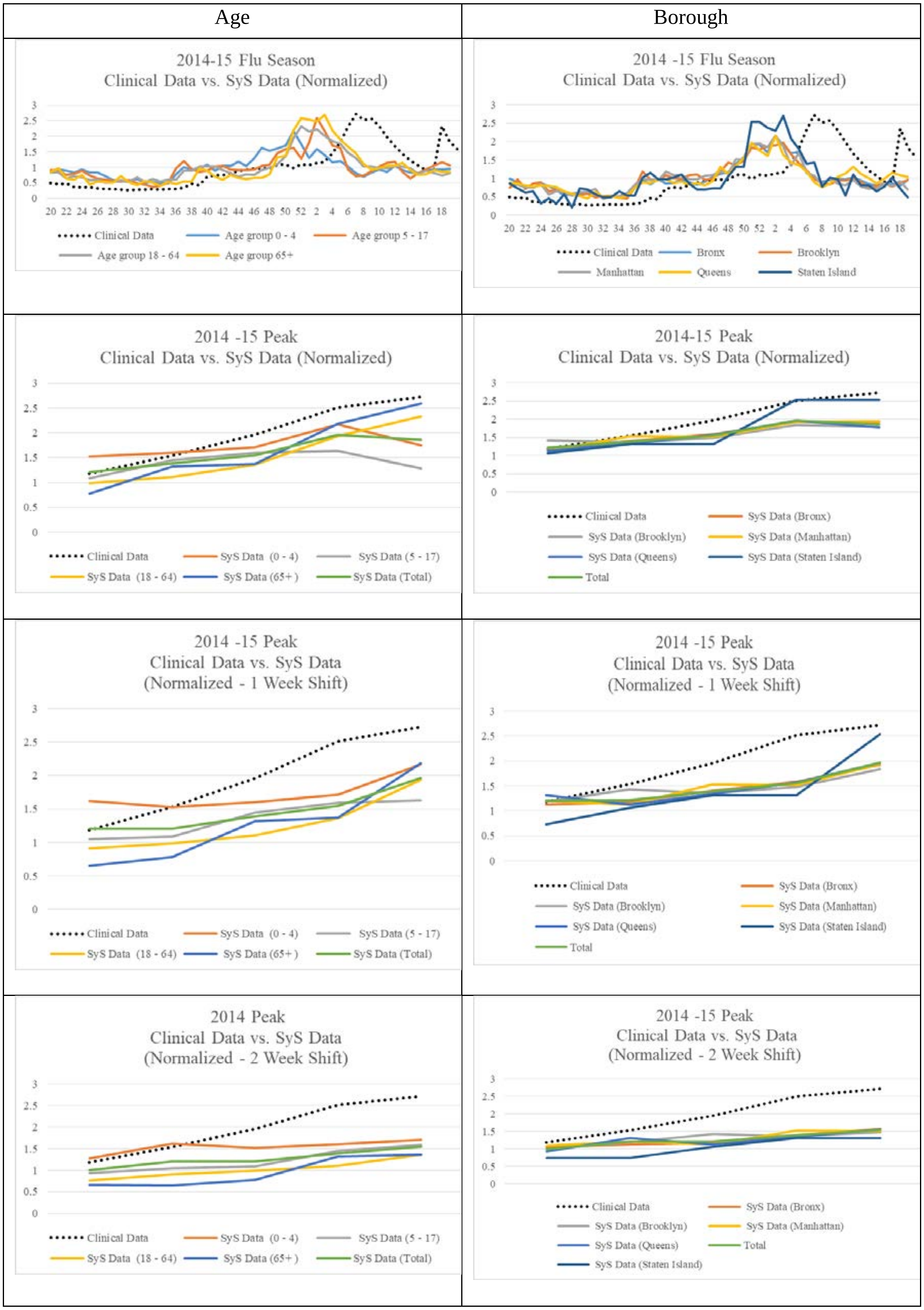


Figure 15 Age and Borough Data from the SyS data subpopulations compared the clinical lab data for the entire flu season as well as various time shits

Correlation

Age

	0-4	5-17	18-64	65+	Total
2 Week	0.954	0.842	0.926	0.647	0.973
1 Week	0.936	0.962	0.957	0.866	0.984
No Shift	0.875	0.994	0.902	0.885	0.943

Summary of age groups with correlation above .8			
	No Shift	1 Week	2 Week
Peak	All – 5-17 (Max)	All – 5-17 (Max)	0-4, 18-64, 5-17

Borough

	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Total
2 Week	0.890	0.924	0.789	0.919	0.841	0.973
1 Week	0.926	0.947	0.962	0.919	0.865	0.984
No Shift	0.974	0.903	0.974	0.885	0.854	0.943

Summary of boroughs with correlation above .8			
	No Shift	1 Week	2 Week
Peak	All – Bronx (Max)	All – Manhattan (Max)	All – Brooklyn (Max)

RMSE

RMSE	0-4	5-17	18-64	65+	Total
Flu Season	0.457	0.425	0.267	0.243	0.344
Peak - 2 Week	0.641	0.833	1.034	1.066	0.817
Peak - 1 Week	0.906	0.911	1.297	1.363	1.039
Peak - In Sync	0.503	0.772	0.463	0.369	0.497
Lowest RMSE	**Group**	**RMSE Value**			
Flu Season	65+	0.243			
Peak - 2 Week	0-4	0.641			
Peak - 1 Week	0-4	0.906			
Peak - In Sync	65+	0.369			

RMSE	Bronx	Brooklyn	Manhattan	Queens	Staten Island	SyS Total
Flu Season	0.329	0.364	0.357	0.380	0.297	0.344
Peak - 2 Week	0.817	0.812	0.806	0.839	1.008	0.817
Peak - 1 Week	1.062	1.041	1.075	1.017	1.308	1.039
Peak – No Shift	0.476	0.571	0.476	0.527	0.324	0.497
Lowest RMSE	**Group**	**RMSE Value**				
Flu Season	Staten Island	0.297				
Peak - 2 Week	Manhattan	0.806				
Peak - 1 Week	Queens	1.017				
Peak – No Shift	Staten Island	0.324				

Figure 16 Summary of both Correlation and RMSE Calculations for SyS and Clinical Data

Figure 16 shows the correlation values for each flu season and the lowest RMSE for each flu season (considering five weeks leading to the peak), respectively.

For this flu season, in all instances measuring RMSE a subpopulation of SyS data has lower RMSE compared to using the entire SyS data set. Conflict between RMSE and correlation calculations were found; for example, with a shift of 2 weeks for the peak week group 65+ has the lowest RMSE, but also the lowest correlation. For borough data, all of the subpopulations performed relatively well and similarly for both RMSE and Correlation at each time shift. For borough data, all data sets are showing high correlation, so, the next section will look at the correlations across all subpopulations to translate these calculations into meaning full results for ModSySIR.

4.2.3 Examining Subpopulation Correlation and Summarizing Results

Overall, there were no obvious patterns identified in the analysis of the correlation values. The data showed that looking for frequency in subpopulations may prove to be more fruitful to identify a pattern in subpopulation behavior. While no clear pattern emerged immediately, an exploration of correlation values at thresholds of 0.8, 0.9, and 0.95 was completed. Results for correlation threshold count are listed in figure 17:

Threshold 0.8	No Shift	1 Week	2 Weeks	Total (of 24)	
Brooklyn	7	4	4	15	
Bronx	7	4	5	16	
Manhattan	7	5	5	17	
Queens	8	5	5	18	Best Case
Staten Island	5	4	4	13	Worst Case

Threshold 0.9	No Shift	1 Week	2 Weeks	Total (of 24)	
Brooklyn	7	1	3	11	
Bronx	6	0	3	9	
Manhattan	6	3	2	11	
Queens	8	2	4	14	Best Case
Staten Island	2	1	2	5	Worst Case

Threshold 0.95	No Shift	1 Week	2 Weeks	Total (of 24)	
Brooklyn	7	0	1	8	Best Case
Bronx	5	0	1	6	
Manhattan	4	0	1	5	
Queens	5	0	3	8	Best Case
Staten Island	2	0	0	2	Worst Case

Figure 17 Frequency of each borough occurring above thresholds (0.8, 0.9, and 0.95) based on peak in the data set

The figures 17 & 18 summarize how many times in all years each subpopulation has a correlation higher that 0.8, 0.9, and 0.95. The goal of this exercise is to find a pattern in frequency and identify a subpopulation with consistent correlation values rather than just focusing on maximum correlation or minimum RMSE values. The borough of Queens seems the most consistent borough in terms of determining the correlation of SyS data for these time periods. Staten Island is obviously a poor choice based on this threshold count and it lacks robust data compared to other subpopulations. Queens and Brooklyn are candidate subpopulation for the ModSySIR model based on frequency and data availability.

0.8	No Shift	1 Week	2 Weeks	Total (of 24)	
0-4	5	1	4	10	
5-17	4	3	2	9	Worst Case
18-64	7	5	5	17	Best Case
65+	6	4	3	13	
0.9	No Shift	1 Week	2 Weeks	Total (of 24)	
0-4	4	1	2	7	
5-17	3	2	2	7	
18-64	5	1	2	8	Best Case
65+	4	2	2	8	
0.95	No Shift	1 Week	2 Weeks	Total (of 24)	
0-4	4	0	0	4	
5-17	2	1	1	4	
18-64	3	0	1	4	
65+	1	1	1	3	

Figure 18 Frequency of each age group occurring above thresholds (0.8, 0.9, and 0.95) based on peak in the data set

The subpopulations 18-64 and 0-4 seems the most consistent age groups in terms of determining the correlation of SyS data for these time periods. 65+ is inconsistent with data availability in the overall SyS data set. Thus 18-64 and 0-4 age groups are also candidate subpopulations for the ModSySIR model.

	Best Case	Worst Case
0.8	18-64	5-17
0.9	18-64, 65+	0-4, 5-17
0.95	0-4, 5-17, 18-64	65+

Summary of threshold counts		
	Best Case	Worst Case
0.8	Queens	Staten Island
0.9	Queens	Staten Island
0.95	Queens, Brooklyn	Staten Island

Lowest Average	65+
Worst Case	65+
Best Case	0-4, 18-64

Lowest Average	Staten Island
Worst Case	Manhattan
Best Case	Brooklyn, Queens

4.3 Epidemic Modeling with Subpopulations and Parameters

Figures 19 shows the basic SIR system of equations with notional parameters. As the infection rate and recovery rate change, the number of individuals in each SIR compartments would also change.

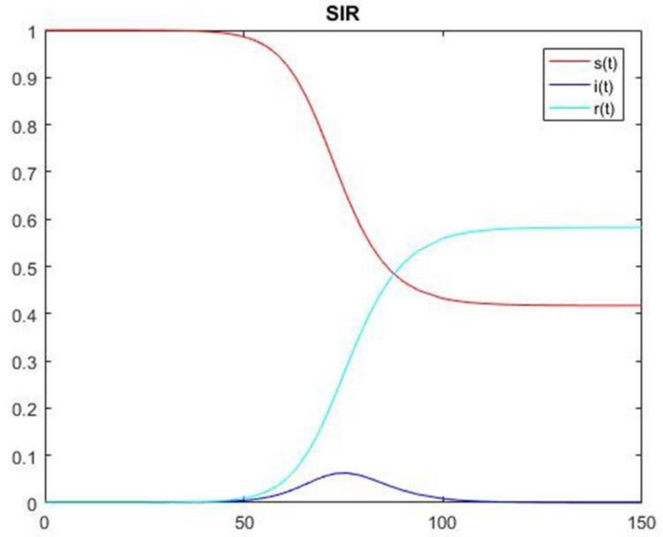


Figure 19 A notional example of the basic SIR model where s(0) ≈ 1, i(0) = 1.2700e-06 r(0) = 0,β=.5, γ=.33. In this case N=1 and compartments S, I, & R will always combine to equal 1; regardless of time, the population will remain unchanged because this is no influx or exit from this closed model.

Initial analysis of SIR showed how the change in parameters and additional compartment changed with increases in infected values and change in peak and locations. The transmission parameter is embedded into the SIR model capture more realistic behavior oh the community. The next step is to include the behavior of subpopulations and analyze the integration of the modification utilizing MATLAB software. The subpopulations are evaluated by time of peak and changes in the "Infected" compartment. Finally, a demonstration of four subpopulation (based on analysis and infection rate found from SyS data and appropriate transmission rate from literature) in the ModSySIR model will be presented. Using real influenza lab data, the MATLAB results will be evaluated for best subpopulation in hopes of real world applicability.

The various stages of data analysis of SyS data and the comparison of SyS data and Lab data presented four different yet representative candidate subpopulations to test: Ages 0-4 & 18-64, and the boroughs of Queens and Brooklyn.

Parameter	Value	Source
Ro=infection rate/recovery	1.3 – Seasonal 1.6 - Pandemic	Coburn, Wagner, & Blower, (2009).
Transmission Coefficient	ρ_i = .5	Ajelli & Litvinova (2017)
Subpopulation	Ages 0-4 & 18-64, and the boroughs of Queens and Brooklyn	NYC Data & Lab Data
Infection Rate (s)	0.153, 0.128, 0.0286, 0.036	NYC Data & Lab Data
Recovery Rate Ro=1.3	0.1179, 0.0982, 0.0220, 0.0283	Extrapolate from Ro
Recovery Rate Ro=1.6	0.0737, 0.0614, 0.0138, 0.0177	Extrapolate from Ro

Figure 20 Parameters for ModSySIR

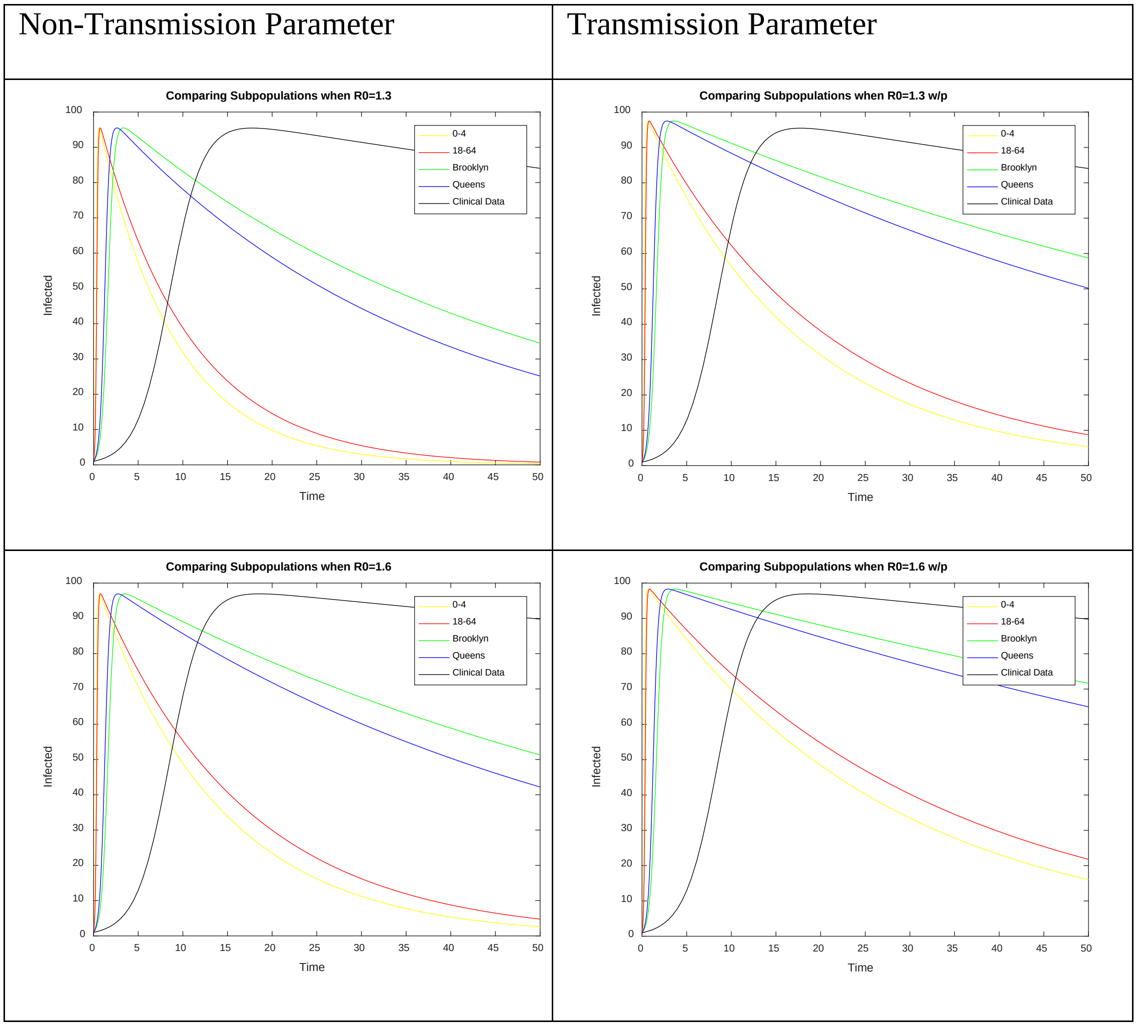
The ModSySIR equation for the infected compartment is:

$$\dot{I}_i = \rho_i \beta S_i I_i - \gamma I_i$$

Based on R_o and the infection rate identify by candidate subpopulations, R_o can be determined. Once the ModSySIR runs through the solver for each R_o, the transmission parameter coefficient is included. Figure 20 holds parameters that inform charts in figure 21, comparing various Ro values with and without the transmission coefficient for the 4 candidate subpopulations. Performance of all of the candidate data sets is graphed and can be found in section 4.4 and implications are discussed in Chapter 5.

4.4 Model Application

In all cases, SyS data from Brooklyn, is closest to clinical data. For age, group 18-64, is the best choice for a representative subpopulation. The reproductive number was varied between known values for seasonal (1.3) and pandemic (1.6) behavior to show how results varied in each situation. The results in Figure 21 highlight the peak time for each subpopulation based on their individual infection rate (which was from SyS data). The associated time of peak values for the varying scenarios and parameters are found in Figure 22.


Non-Transmission Parameter

Transmission Parameter

Comparing Subpopulations when R0=1.3

Comparing Subpopulations when R0=1.3 w/p

Comparing Subpopulations when R0=1.6

Comparing Subpopulations when R0=1.6 w/p

Figure 21 Candidate representative subpopulations in ModSySIR compared to baseline clinical lab data. In this case N=100, rather than N=1.

Time of Peak

	R_o **=1.3**	R_o **=1.3+p**	R_o **=1.6**	R_o **=1.6+p**
0-4	0.628	0.668	0.649	0.697
18-64	0.753	0.802	0.778	0.837
Brooklyn	**3.361**	**3.579**	**3.474**	**3.735**
Queens	2.624	2.788	2.707	2.911
Lab Data	*17.584*	*17.584*	*18.399*	*18.399*

Figure 22 Summary of time peak occurrences in ModSySIR with varying R_0 and inclusion of transmission parameter

In the 4 different scenarios, there is similar behavior as Ro increases and transmission coefficient is added because these changes occur at the same rates for each subpopulation. So in all cases, Brooklyn is the best representative subpopulation. The potential risk of using the wrong subpopulation is further discussed in Chapter 5.

Based on the analysis the subpopulation and associated parameters for model application are:

- Best Subpopulation: Brooklyn
- Appropriate Parameters: Infection Rate of 0.0286, not including transmission coefficient

Chapter 5: Discussion

5.1 Results Summary

The entire SyS data analysis showed normal seasonal flu behavior as well as an unusual increase in encounters for the 2009-2010 flu season. It's valuable to have this variety of activity when looking to model both epidemics and both pandemics. Unfortunately, the analysis results initially showed no clear leader for the best subpopulation. There were inconsistent results for subpopulation indicators for each year. The illustration of the 2014-15 flu season in Chapter 4 was key to developing a process for a public health official to identify and rank subpopulations based on data analysis and availability. For the process to be efficient, the data analysis would require automation to translate subpopulation data and calculate appropriate statistics and parameters. Overall, there were no obvious patterns identified in the initial analysis of the correlation values. In an effort to pinpoint patterns within the data comparison process, two ways to measure likeness in data have been presented. Various means helped to identify the best representative subpopulation that may have otherwise been ruled out by utilizing just one error calculation. Taking a look from a different perspective helped to identify trend in the correlation and RMSE calculations. The most useful insights came from peak analysis and comparison of behavior among various subpopulations and data sets. Focusing on just one flu season to identify representative parameters is not ideal and does not necessarily translate to the entire data set. Regardless of representation, robust data is better than sparse data. A subpopulation that only represents small portion of a population such as Staten Island or the 65+ groups wouldn't not be the right choice for the ModSySIR model.

Table 5 shows the infection rate and recovery rate for the four candidate subpopulation, clinical data, and the age group of 65+

	0-4	18-64	Brooklyn	Queens	Clinical Data	65+
Infection Rate	0.1532	0.1277	0.0286	0.0367	0.0054	0.1914
Recovery Rate; Ro=1.3	0.1179	0.0982	0.0220	0.0283	0.0042	0.1472
Recovery Rate; Ro =1.6	0.0737	0.0614	0.0138	0.0177	0.0026	0.0920

For example, 65+ age group was run through the same process in MATLAB to compare the peak time in the ModSySIR model similar to the candidate subpopulations. 65+ age group is only 3% of the total data set; thus was initially discounted from being a representative subpopulation. Table 6 shows the associated parameters of this age group in ModSySIR. There is risk is choosing the wrong subpopulation to represent the overall population.

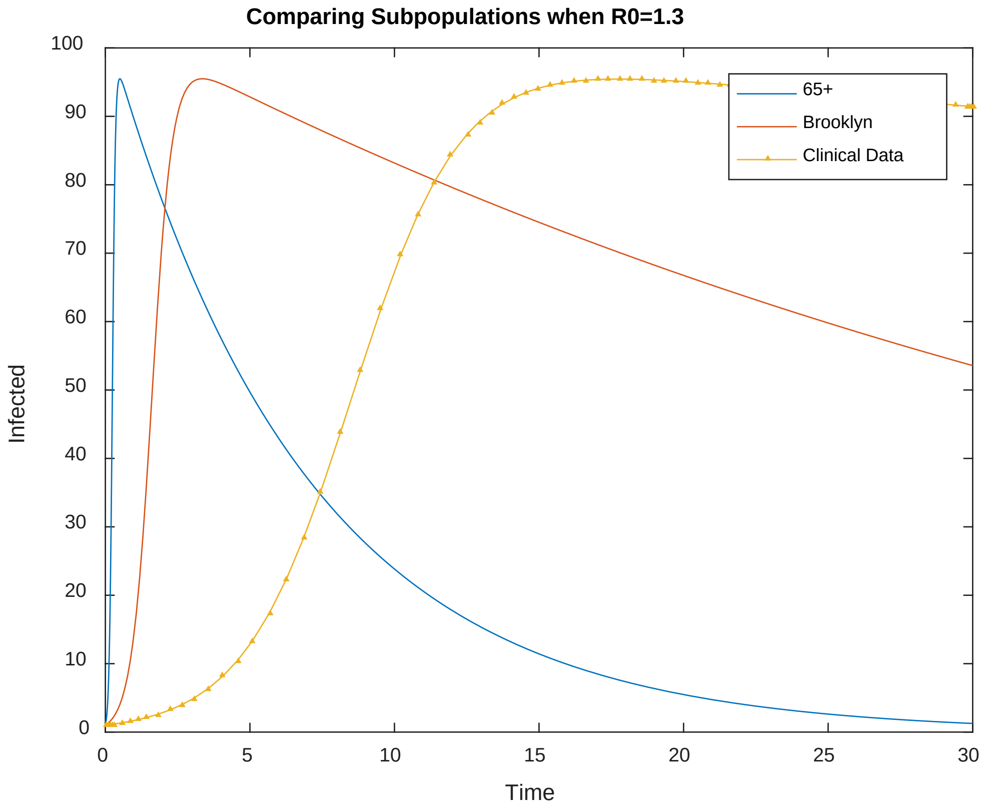


Figure 23 The best candidate subpopulation, Brooklyn, and an unsuitable subpopulation, 65 +, compared to the Clinical data.

In Figure 23, the time of peak occurrence for the clinical data, 17.54, is closer to the Brooklyn peak time of 3.36, rather than the 65+ peak time of .5.

5.2 Research Questions

This research effort started with four basic questions that this analysis process has worked to provide answers for. This section will highlight the questions and also the process and answer, if determined. Based on these answers, the next section dives into some insights and a potential implementation plan.

Which subpopulation's behavior best represents the overall encounters of the SyS data set? Various subpopulations have high correlation and low RMSE for each year. Looking from year to year there was no subpopulation that was always or most often the best representative of the dataset. The most valuable information came from flu season and peak data analysis, rather than looking at the entire calendar year for likeness when comparing data. There was no clear indicator from the analysis of 8 flu seasons which subpopulation was most representative; therefore further analysis was needed. A use case of the 2014-15 flu season was explored to better understand how to narrow down best and worst case scenarios when selecting a subgroup within the data.

Using the most representative age group and location can analysis determine which subpopulation is more likely to miss an increase in overall ILI activity? In an effort to still determine the most representative group, there were four candidate subpopulations identified: Ages 0-4 & 18-64, and the boroughs of Queens and Brooklyn. Of those groups, parameters determined from age group 0-4 provide the poorest match the Clinical data within the ModSySIR model. This is illustrated graphically in figure 21 and a numerical comparison of each group's parameters in figure 22 also shows how far this value is from the clinical data set.

Which, if any, subpopulation(s) show leading indicators of an impending outbreak? Which demographic category is better to use (an age group or location) to represent the population? There is no clear cut answer. Inferences could be made from the correlation threshold counts regarding the frequency of correlation and RMSE values using data from figures 16 and 17. Overall, Brooklyn was the best subpopulation; therefore, location is the best demographic in this instance. It also represented a fair amount of the population so it was likely to capture more fidelity in the behavior of the overall population.

Can a model utilizing subpopulation data and/or behavior provide representative and predictive information for the entire population? The ModSySIR model was more appropriate for data fitting of subpopulation parameters rather than prediction. Predictive uses and extensions of these methods are explored in Chapter 6 Section 5.

5.3 Implications and Insights

Overall the data analysis process was a useful way to investigate the available dataset, gather general knowledge and better understand community and subpopulation characteristics. The process presented many ideas for potential applications in limited resource communities and countries. For implementation of the data analysis, communities need to choose to survey this process and identify appropriate techniques based on quality of data and availability of resources. For example, NYC SyS allowed for the collection of over 600,000 encounters over an eight year period. This comprehensive reporting infrastructure is not likely established in resource limited communities. In such communities, health departments should start with a survey about social contacts from a sample group of constituents to determine how to potential disease modeling based on social behavior. ModSySIR considers a predefined subpopulation transmission coefficient. The use of an SIR model rather than another epidemic model has 3 immediate benefits: 1) Requires limited data, 2) Simple implementation of system of equations, and 3) Demonstrated and substantiated modifications.

Based on lessons learned from this research, an overview of a potential implementation guide for a resource limited community is listed below:

1. Data Availability and Analysis
    a. Public health officials should assess available data based on quantity and population coverage. Next step is to assess data activity and/or behavior and fluctuations in data. Then the peaks of subsets and the overall data set should be recorded and potential techniques to assess relationships among the data subsets, such as correlation, should be evaluated. Finally the results are evaluated to determine if there are any patterns in the data.
2. Community Behavior
    a. Learn about the health care seeking behavior of the community of interest and identify any remarkable characteristics, such as an extremely susceptible population. Community characteristics such as socio-economic factors, and demographics, cultural norms, and population density are relevant when studying the community. These factors can help to determine heterogeneous or homogeneous social mixing. Then an appropriate social contact matrix or transmission parameter is chosen.

3. Determine Parameters
    a. Determine the reproductive number in which the parameters will be extrapolated, considering calculations based on R_o and infection rate of subpopulation of interest. The infection rate is determined from the cumulative encounters leading to the peak divided by the overall number of encounters or the population. Identify the baseline data for comparison to observed data or training data.
4. Apply Parameters to ModSySIR
    a. Apply parameters predetermined in step 3 to identify the time of the peak occurrence from each subpopulation of interest.
5. Review Results and Supplement data
    a. Recognize other sources of data for syndrome of interest and use supplemental surveillance data as needed as an additional resource. Determine best set of parameters to compare to clinical data set.
6. Make decision and/or inform policy
    a. Public health officials have many decisions to make and limited time when presented with a syndrome potential harmful but also time sensitive. Based on data analysis and parameter selection. Public health officials need to work to form a response to prepare for and additional outbreak. Determine reporting, data use, and/or plan of action.

Public Health Officials have a lot of responsibilities and important decisions to make to ensure their community is well monitored and prepared for public health events. This research provides a clear and effective method for data analysis. In turn, public health officials are provided with a method to identify appropriate model parameters and monitored data subset. In fact, supplemental data provides addition information for decision support.

Chapter 6: Research Summary

6.1 Process

Overall, this research process combined multiple topics including ILI, varying data sources and stratification, epidemic modelling, and SyS data. Robust data analysis was used to characterize data and subpopulation behavior. This analysis focused on the flu season and peak data for each year; moreover, various statistics and measures of likeness were calculated to understand similarities and detect differences in the data, both SyS and clinical. Parameters for subpopulations were compared to clinical data. Based on this process, an implementation guide for public health official was identified.

6.2 Results

Based on the illustration from the 2014-15 flu season, this process identified the best candidate subpopulation as Brooklyn. Also, there are inadequacies with using a subpopulation with limited data, recognized by examining the 65+ data set. With or without the inclusion of the transmission parameter, Brooklyn was the best population to select parameters for ModSySIR. This insight is unique to this data set and will likely be different with new data set and/or community. Similarly, the transmission parameter could be play a role in identifying appropriate parameters in a candidate subpopulation with a different data source.

6.3 Contribution and Implementation

This research has identified representative subpopulations for identification of ILI peaks through an in-depth data analysis process. Resource limited health departments can follow a similar data analysis process, based on data availability and robustness, to characterize their population and identify a subset to monitor for potential outbreaks. Overall, this customizable SIR model incorporates preexisting community data to identify the best SyS subpopulation to monitor for indication of impending disease behavior.

Public health officials in resource limited communities need to scrutinize the use time, money, and personnel for public health surveillance. If applying this research methodology, public health officials need to identify their starting point based on available data. One option is to follow the implementation process through data analysis and narrow down candidate subpopulations. Next, they would compare their subpopulation finding with clinical data if available, otherwise, comparison of behavior similarities with the entire data set is appropriate; however, if initially, time or data do not allow the robust analysis then public health officials should look to other country or communities that mirror the socio economic characteristics, density, demographics, and health care seeking behaviors of their own community. This can at least assist with an initial selection of a population to monitor and determine near term decisions such as intervention strategies. As resources become more constrained, public health officials need to be even more careful of the subpopulation they target. In this case consider, varying which subpopulation is monitored for each epidemic to better understand community activity by comparing that monitored subpopulation to confirmed data.

The process of separating a total population into subpopulations for the purposes of this survey need to be consistent, so pre-established groupings such as those in the NYC data or in other population stratified models would be appropriate. For heterogeneous mixing, establishing a contact matrix based on self-reported community behavior including the proportion of the population that each subpopulation represents. In the

ModSySIR model, initial conditions, such as size, for the susceptible population is based on community population data, infected population relies on SyS data to determine infection rate, and recovered population would be zero. The contact matrix is inclusive of all subpopulations and quantifies contacts based on survey results. Output of the SIR model is time series data. For each population, analyze notable changes such as beginning and duration of an increase or decrease. Changes in population density call for a review of contact patterns as well. Charts from the ModSySIR consistently show how Brooklyn maps to the clinical laboratory reports. In this case, next steps would include continuous review of incoming data now with assigned parameters based on the identified data and community implications.

6.4 Limitations

The main limitation of this research is the type of epidemic model used. SIR is a basic model, chosen for the historical evidence of parameter manipulation, as opposed to an option such as agent based modeling that would consider heterogeneous individual behavior or even machine learning. The transmission parameter was chosen in lieu of a community specific contact matrix. A resource limited health department should be empowered to choose the right level of complexity in which they enact this approach. In addition, the SIR model chosen relies on a constant population; thus, fluctuations in the population could be included such as models with mortality parameters like those presented in the literature review. Realistically, unless the monitored syndrome is a pandemic or the infection lingers over an extended period, this is negligible for this approach. The data available was very useful for analysis and application; however, it is from an urban setting in the United States and the volume and quality of the data does not likely mimic the data available to a resource limited community. More data from various environments is needed for endorsement. This circles back to the original motivation of this effort: the lack of research in resource limited communities and various potential application opportunities for those areas.

Realistically, SyS data is intended to be an early warning system for syndromes of interest; however, some groups are more prone to go to seek care at unmonitored locations so they may not be well represented in SyS data. Also, different groups may show better indicators for different syndromes. Public health official need a plan to determine a similar population to their community or country if they do not have the available resources to set a baseline for data comparison. There are a lot of factors to consider when choosing a common data set for an instance like the one previously mentioned. In fact, this research was possible because of a large publically available data set. There is additional research needed to identify what amount of data is appropriate to consider a sample subpopulation size appropriate considering the methods in place for this effort. Pre-sorted and categorized groups within the NYC SyS data were available, but in other cases where the de-identified data is not categorized other means of data organization and analysis will need to occur based on data detail and availability.

6.5 Extension

Section 2.3 briefly mentioned capabilities of Machine Learning (ML). ML techniques can create generalized models for similar communities to utilize which could prove useful when a public health official needs to use a data or parameters from another community (Weins and Shenoy, 2018). These algorithms have been used for data set shifts, similar to those seen in this research, to recognize predictive utility, using various models and integrating multiple data sources (Xu et al., 2017, Santillana et al., 2015). Efforts have been made to relate information in closely related environments to help with prediction of another data set (Quionero-Candela et al., 2009). In fact, ML can serve as a decision support system for public health officials. A study by Brownstein et al. (2017) reviewed predictions from 1, 2, 3, & 4 weeks in advance from 6 different countries. Their geographically diverse research could provide discernment for a public health official to match their country or community to the right model parameters. The implementation from Chapter 5 could also be extended to multiple data sources depending upon availability.

Understanding the risk of inaccurate prediction and how improper data partitioning can affect ML model performance is important in analysis and assessment. Even with these potential extensions, the models can still be evaluated with Pearson's correlation coefficient and RMSE (Brownstein et al., 2017 and Santillana et al., 2015). With ML, large data sets are need so further research on the minimum data set sample size based on population and model choice is required for further application. In general, ML could potentially be an effective way to evaluate total flu season as opposed to just focusing on peak encounters. Translating these insights to identification of relative risk of each subpopulation with automation and algorithm evolution could prove to be of great benefit to public health decision makers.

6.6 Conclusion

This research motivation was to explore opportunities to assist resource limited population by enhancing their Syndromic Surveillance (SyS) systems. Organizing the collected SyS data into predetermined subpopulations (separated by population characteristics such as age or location) allowed evaluation of models, associated parameters, and real data correlate to disease progression. Modifications of the SIR model were assessed and adapted for ModSySIR. The research provides a valuable modeling option, dependent on limited data and familiar mathematical methods, to resource limited communities. Fumanelli at al. (2012) believed that models like ModSySIR that work to predict behavior for public health entities can be extended to “every country with socio-demographic data”. Potential implementation of ModSySIR model will better equip and information public health entities in such communities.

Appendix A

Borough Correlation Data (Flu Season):

2008-09	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.975993	0.919949	0.895241
	Brooklyn	0.94175	0.981735	0.935495
	Manhattan	0.99124	0.964674	0.922132
	Queens	0.964817	0.923878	0.882612
	Staten Island	0.718806	0.744569	0.766914

2009-10	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.992924	0.82757	0.575094
	Brooklyn	0.987426	0.769031	0.554494
	Manhattan	0.976949	0.901376	0.780051
	Queens	0.990132	0.792363	0.577245
	Staten Island	0.943523	0.785813	0.625715

2010-11	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.99245	0.91058	0.784891
	Brooklyn	0.988909	0.923159	0.82405
	Manhattan	0.97975	0.861207	0.692211
	Queens	0.994947	0.928851	0.806058
	Staten Island	0.815105	0.690697	0.52528

2011-12	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.896089	0.864167	0.685946
	Brooklyn	0.875307	0.880121	0.764839
	Manhattan	0.873926	0.795108	0.677833
	Queens	0.95172	0.812008	0.62551
	Staten Island	0.812108	0.857358	0.695619

2012-13	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.995609	0.89957	0.768943
	Brooklyn	0.989471	0.889169	0.707765
	Manhattan	0.988943	0.882357	0.693784
	Queens	0.985716	0.942003	0.840637
	Staten Island	0.898345	0.946778	0.826895

2013-14	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.984906	0.858198	0.695999
	Brooklyn	0.964956	0.841871	0.799122
	Manhattan	0.965066	0.856082	0.735339
	Queens	0.985604	0.892341	0.766692
	Staten Island	0.81271	0.731929	0.699167

2014-15	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.98702	0.878444	0.778259
	Brooklyn	0.970479	0.885044	0.753402
	Manhattan	0.965072	0.851845	0.768754
	Queens	0.953558	0.89	0.780513
	Staten Island	0.912653	0.740504	0.622744

2015-16	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.973288	0.904192	0.833373
	Brooklyn	0.980562	0.911577	0.799418
	Manhattan	0.940263	0.849359	0.762507
	Queens	0.921444	0.890939	0.771292
	Staten Island	0.809338	0.716138	0.578174

Summary of Maximum Correlations Values for Each Flu Season

	No Shift	1 Week	2 Week
2008-09	Brooklyn	Brooklyn	Brooklyn
2009-10	Bronx	Manhattan	Brooklyn
2010-11	Bronx	Brooklyn	Manhattan
2011-12	Bronx	Queens	Brooklyn
2012-13	Manhattan	Queens	Bronx
2013-14	Manhattan	Brooklyn	Queens
2014-15	Manhattan	Brooklyn	N/A
2015-16	Manhattan	Queens	Queens

Borough Correlation Data (Peaks Only)

2008-09	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.638789	0.785804	0.638789
	Brooklyn	0.96326	0.952547	0.96326
	Manhattan	0.913125	0.851296	0.913125
	Queens	0.952421	0.825054	0.952421
	Staten Island	-0.9195	0.273372	-0.9195

2009-10	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.96843	0.982964	0.96843
	Brooklyn	0.98135	0.895306	0.98135
	Manhattan	0.993093	0.909277	0.993093
	Queens	0.974416	0.949959	0.974416
	Staten Island	0.980274	0.987092	0.980274

2010-11	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.984798	0.815459	0.984798
	Brooklyn	0.981327	0.810681	0.981327
	Manhattan	0.955145	0.822653	0.955145
	Queens	0.994791	0.874611	0.994791
	Staten Island	0.465479	0.43959	0.465479

2011-12	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.764508	0.596177	0.764508
	Brooklyn	0.005913	0.090474	0.005913
	Manhattan	0.81206	0.36582	0.81206
	Queens	0.968074	0.532516	0.968074
	Staten Island	0.576349	0.462935	0.576349

2012-13	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.982747	0.736277	0.982747
	Brooklyn	0.975137	0.940303	0.975137
	Manhattan	0.979418	0.944979	0.979418
	Queens	0.963966	0.800305	0.963966
	Staten Island	0.556172	0.898519	0.556172

2013-14	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.991778	0.938017	0.991778
	Brooklyn	0.997556	0.865951	0.997556
	Manhattan	0.992232	0.891141	0.992232
	Queens	0.996998	0.918134	0.996998
	Staten Island	0.7786	0.824227	0.7786

2014-15	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.989564	0.3917	0.989564
	Brooklyn	0.849384	0.539204	0.849384
	Manhattan	0.950665	0.197331	0.950665
	Queens	0.971802	0.241386	0.971802
	Staten Island	0.832456	0.243049	0.832456

2015-16	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Bronx	0.989564	0.3917	0.989564
	Brooklyn	0.849384	0.539204	0.849384
	Manhattan	0.950665	0.197331	0.950665
	Queens	0.971802	0.241386	0.971802
	Staten Island	0.832456	0.243049	0.832456

Summary of Maximum Correlation Values for Each Flu Season

	No Shift	1 Week	2 Week
2008-09	Brooklyn	Brooklyn	Brooklyn
2009-10	Manhattan	Staten Island	Manhattan
2010-11	Queens	Queens	Queens
2011-12	Queens	N/A	Queens
2012-13	Bronx	Manhattan	Manhattan
2013-14	Brooklyn	Bronx	Brooklyn
2014-15	Bronx	N/A	Bronx
2015-16	Bronx	N/A	Bronx

Age Correlation (Peaks Only)

	Correlation w/ Time Shift	No Shift	1 Week	2 Week
2008-09	Age 0-4	0.713887	0.827819	-0.62489
	Age 5-17	0.99082	0.981365	0.985394
	Age 18-64	0.954437	0.876046	0.660392
	Age 65+	0.797622	0.076519	-0.87939

	Correlation w/ Time Shift	No Shift	1 Week	2 Week
2009-10	Age 0-4	0.975205	0.939797	0.891954
	Age 5-17	0.969099	0.908791	0.957932
	Age 18-64	0.81796	0.963413	0.906543
	Age 65+	0.807448	0.878459	0.400263

	Correlation w/ Time Shift	No Shift	1 Week	2 Week
2010-11	Age 0-4	0.976963	0.697406	0.817526
	Age 5-17	0.866717	0.942797	0.812062
	Age 18-64	0.958964	0.806216	0.870491
	Age 65+	0.971428	0.792676	0.812024

	Correlation w/ Time Shift	No Shift	1 Week	2 Week
2011-12	Age 0-4	0.906066	0.483174	0.80268
	Age 5-17	0.135851	0.336748	0.990515
	Age 18-64	0.903556	0.147459	0.752025
	Age 65+	0.465	0.776808	-0.75256

	Correlation w/ Time Shift	No Shift	1 Week	2 Week
2012-13	Age 0-4	0.71509	0.634694	0.986809
	Age 5-17	0.658693	0.23159	0.761404
	Age 18-64	0.974998	0.908468	0.939271
	Age 65+	0.9722	0.954713	0.892497

	Correlation w/ Time Shift	No Shift	1 Week	2 Week
2013-14	Age 0-4	0.954656	0.726919	0.294634
	Age 5-17	0.99504	0.966572	0.720058
	Age 18-64	0.990192	0.885566	0.686789
	Age 65+	0.970012	0.83906	0.890641

2014-15	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.228067	-0.25342	0.213173
	Age 5-17	0.486906	0.914036	0.267177
	Age 18-64	0.789008	0.38481	0.50351
	Age 65+	0.762689	0.384371	0.560252

2015-16	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.697113	0.337522	0.866142
	Age 5-17	0.783526	0.455658	0.197975
	Age 18-64	0.924634	0.670216	0.979192
	Age 65+	0.922823	0.857199	0.979746

Summary of Maximum Correlation Values for Peaks

	No Shift	1 Week	2 Week
2008-09	Age 5-17	Age 5-17	Age 5-17
2009-10	Age 0-4	Age 18-64	Age 5-17
2010-11	Age 0-4	Age 5-17	Age 18-64
2011-12	Age 0-4	Age 65+	Age 5-17
2012-13	Age 18-64	Age 65+	Age 0-4
2013-14	Age 18-64	Age 5-17	Age 65+
2014-15	Age 18-64	Age 5-17	N/A
2015-16	Age 18-64	Age 65+	Age 65+

Age Correlation (Flu Season)

2008-09	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.681371	0.706295	0.692439
	Age 5-17	0.930282	0.912431	0.901384
	Age 18-64	0.946303	0.834378	0.800195
	Age 65+	0.871166	0.538373	0.381249

2009-10	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.970577	0.681603	0.378465
	Age 5-17	0.922744	0.84833	0.750136
	Age 18-64	0.969054	0.895045	0.793085
	Age 65+	0.812327	0.307578	0.165394

2010-11	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.945159	0.921146	0.875495
	Age 5-17	0.895498	0.924275	0.832696
	Age 18-64	0.931315	0.726761	0.528684
	Age 65+	0.876303	0.742687	0.548933

2011-12	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.907618	0.924126	0.742176
	Age 5-17	0.706358	0.811156	0.791578
	Age 18-64	0.467554	0.367564	0.20507
	Age 65+	0.243192	0.110297	-0.16065

2012-13	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.929994	0.942889	0.930961
	Age 5-17	0.937332	0.815614	0.673157
	Age 18-64	0.979326	0.882646	0.693728
	Age 65+	0.964281	0.873311	0.663031

2013-14	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.408458	0.639067	0.754455
	Age 5-17	0.846129	0.65579	0.566646
	Age 18-64	0.910442	0.706292	0.508692
	Age 65+	0.830669	0.579578	0.433737

2014-15	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.80345	0.765789	0.852611
	Age 5-17	0.899294	0.760661	0.55419
	Age 18-64	0.92277	0.788698	0.653627
	Age 65+	0.900844	0.804329	0.66308

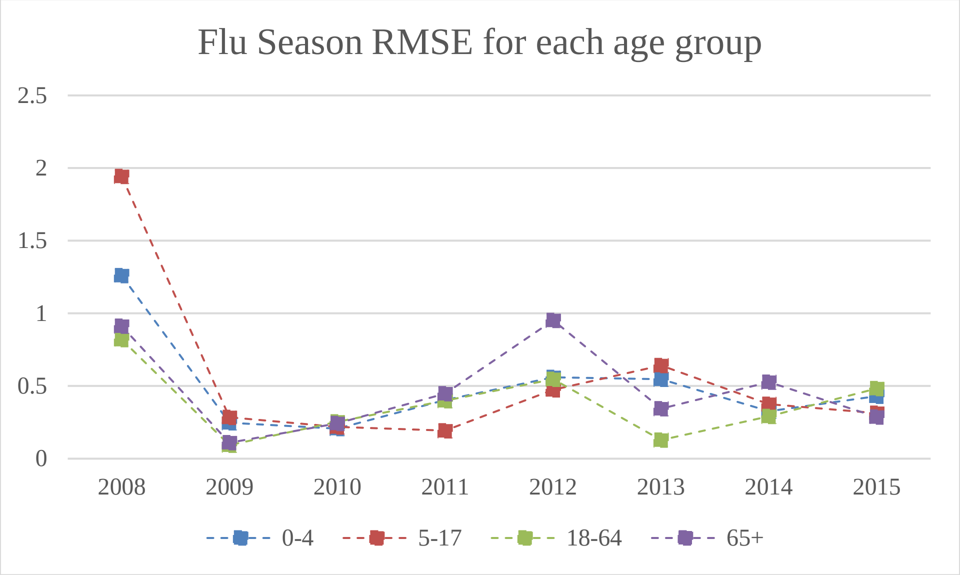
2015-16	Correlation w/ Time Shift	No Shift	1 Week	2 Week
	Age 0-4	0.627002	0.750891	0.66285
	Age 5-17	0.898281	0.916375	0.892154
	Age 18-64	0.898642	0.722919	0.641698
	Age 65+	0.82834	0.434789	0.212739

Summary of Maximum Correlation Values for Each Flu Season

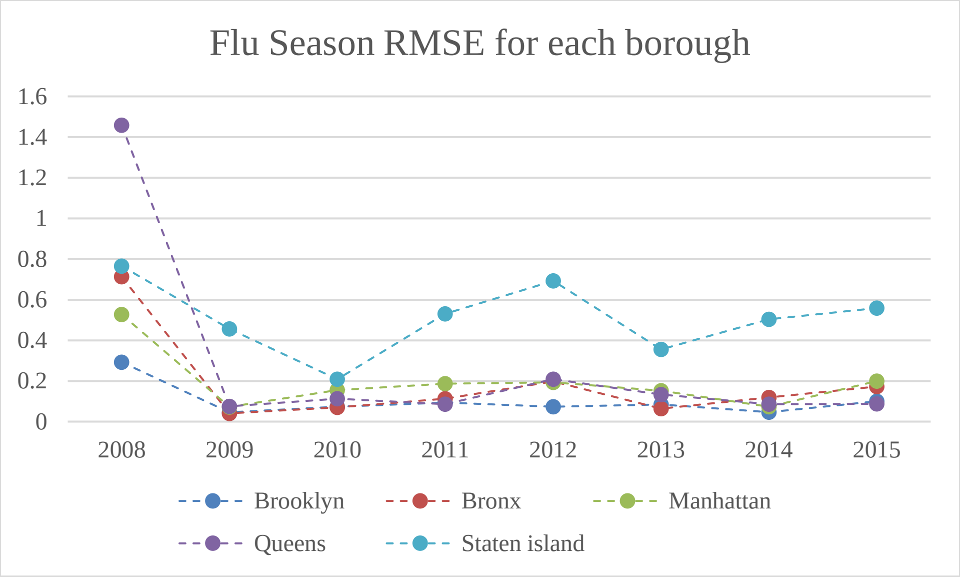
	No Shift	1 Week	2 Week
2008-09	Age 18-64	Age 5-17	Age 5-17
2009-10	Age 0-4	Age 18-64	Age 18-64
2010-11	Age 0-4	Age 5-17	Age 0-4
2011-12	Age 0-4	Age 0-4	Age 5-17
2012-13	Age 18-64	Age 0-4	Age 0-4
2013-14	Age 18-64	Age 18-64	Age 0-4
2014-15	Age 18-64	Age 65+	Age 0-4
2015-16	Age 18-64	Age 5-17	Age 5-17

RMSE for SyS Data

	0-4	5-17	18-64	65+
2008	1.25991	1.943848	0.820426	0.912228
2009	0.248016	0.284478	0.093436	0.109786
2010	0.20662	0.218876	0.253773	0.243067
2011	0.403839	0.19336	0.399335	0.448822
2012	0.55984	0.474596	0.545936	0.952226
2013	0.54652	0.642287	0.128029	0.343975
2014	0.327511	0.375648	0.292092	0.526086
2015	0.429524	0.313738	0.482657	0.286403
Average	0.497722	0.555854	**0.37696**	0.477824
	Lowest Average		18-64	

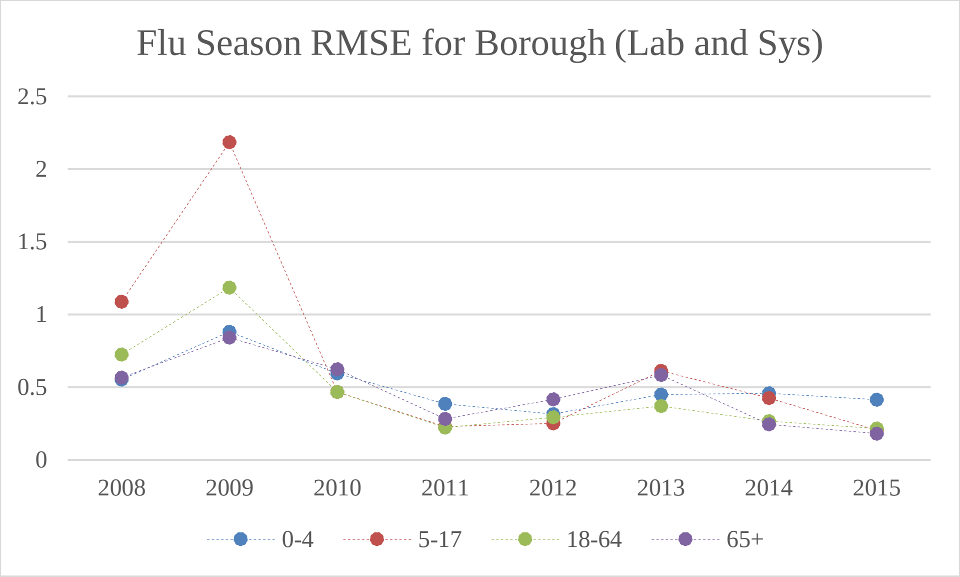


	Brooklyn	Bronx	Manhattan	Queens	Staten Island
2008	0.292652	0.712665	0.526962	1.458669	0.765259
2009	0.045572	0.040689	0.072347	0.075168	0.456072
2010	0.073691	0.069919	0.154571	0.112193	0.208382
2011	0.093926	0.11312	0.187272	0.085917	0.530508
2012	0.073347	0.196865	0.193275	0.209095	0.69196
2013	0.084845	0.06296	0.15267	0.13307	0.355795
2014	0.046989	0.118536	0.073015	0.08544	0.503664
2015	0.099338	0.172927	0.199663	0.087371	0.557986
Average	**0.101295**	0.18596	0.194972	0.280865	0.508703
	Lowest Average		Brooklyn		

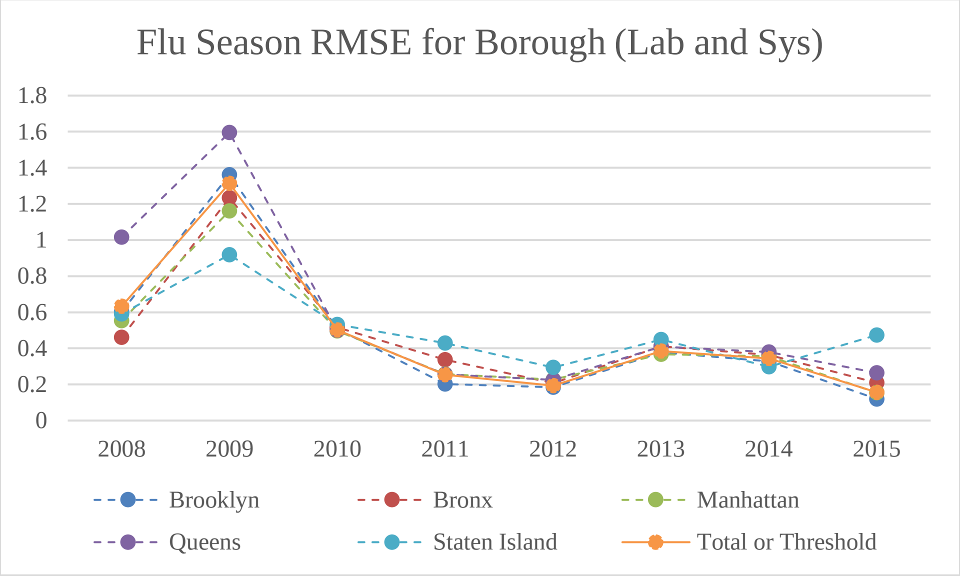


RMSE for Lab and SyS Data

	0-4	5-17	18-64	65+
2008	0.553828	1.086936	0.723791	0.566283
2009	0.881248	2.184042	1.184691	0.840239
2010	0.594653	0.466154	0.467483	0.622263
2011	0.38454	0.227964	0.222633	0.281757
2012	0.314736	0.250564	0.292093	0.416101
2013	0.449264	0.612415	0.370114	0.584654
2014	0.457437	0.425175	0.267332	0.243427
2015	0.414119	0.203784	0.215713	0.181051
Average	**0.506228**	0.682129	**0.467981**	**0.466972**



	Brooklyn	Bronx	Manhattan	Queens	Staten Island
2008	0.602467	0.461294	0.55352	1.016103	0.591126
2009	1.361252	1.233391	1.16106	1.59502	0.918758
2010	0.509401	0.515817	0.496277	0.500521	0.531079
2011	0.202075	0.336501	0.256655	0.25515	0.429463
2012	0.184853	0.207473	0.226961	0.225489	0.294059
2013	0.378483	0.411971	0.367545	0.409099	0.448733
2014	0.328693	0.363538	0.357142	0.3797	0.297258
2015	0.119787	0.208928	0.154377	0.263621	0.473888
Average	**0.460876**	**0.467364**	**0.446692**	0.580588	0.498046



Lab vs. SyS Data Correlation Summary

	2008-09	**2009-10**	**2010-11**	**2011-12**
2 Week	5-17,	0-4, 18-64	No trend	0-4
1 Week	5-17, 18-64	18-64, 0-4	No trend	0-4
No Shift	5-17, 18-64	0-4	No trend	No trend
	2012-13	**2013-14**	**2008-09**	**2015-16**
2 Week	All, 0-4 (max)	5-17, 18-64	0-4, 18-64, 5-17	5-17
1 Week	All 0-4 (max)	5-17, 18-64	All - 5-17 (max)	18-64
No Shift	All, 5-17 (max)	5-17, 18-64	All 5-17 (max)	18-64, 0-4, 65+

	2008-09	**2009-10**	**2010-11**	**2011-12**
2 Week	No trend	Staten Island, Bronx, Manhattan, Brooklyn	No trend	No trend
1 Week	Brooklyn, Manhattan	Manhattan, Bronx	No trend	Bronx
No Shift	Brooklyn, Queens, Manhattan	Manhattan, Staten Island	No trend	No trend
	2012-13	**2013-14**	**2008-09**	**2015-16**
2 Week	All - Queens (max)	Brooklyn, Queens	All - Brooklyn (max)	Brooklyn, Bronx
1 Week	All - Brooklyn (max)	Manhattan, Bronx, Staten Island	All - Manhattan (max)	Brooklyn
No Shift	All - Manhattan (max)	Manhattan, Queens, Staten Island	All - Bronx (max)	All - Manhattan (max)

Data Analysis Process

The following sections step through the data analysis process. The primary phases include the organization, normalization, correlation, and characterization of subpopulation data.

1. Organize the data
    - a. Export data with appropriate subpopulation parameters (age or borough)
    - b. Separate data into appropriate subgroup (ex. 0-4, 5-17 or Manhattan, Queens, etc.)
    - c. Separate subgroup data into 1 year increments
    - d. Aggregate data every 7 days
    - e. Find basic descriptive and summary statistics for each data set
    - f. Identify weeks with maximum and minimum counts (to identify multiple peaks and periods of low activity)
2. Normalize the data:
    - a. Find the average 7 day value for the subgroup
    - b. Divide all 7 day values by the subgroup average value
    - c. Find basic descriptive and summary statistics for each data set
3. Correlate data: Option 1 – Compare change in weekly count (week to week slope)
    - a. Find the difference between each weekly count (ex. 1-2, 2-3, 3-4, etc.)
    - b. Compare the weekly changes of each subgroup to the total population by calculating Pearson’s correlation coefficient (ex 0-4 vs. All, queens vs. All, etc.)
    - c. Repeat this process moving the total population back 1 week and 2 weeks.

4. Correlate data: Option 2 – Compare cross correlation
    a. Compare the week data of each subgroup to the total population using the correlation worksheet in Excel (ex 0-4 vs. All, queens vs. All, etc.)
    b. Utilize shift table to record correlation when total population lags 1 week and 2 weeks.
    c. Identify time shift with highest correlation
5. Identify representative age and borough subgroups
    a. Find one age group and one borough with the highest correlation (without lag)
    b. Combine the age group and borough data (new data set) to determine correlation coefficient
    c. Repeat process for 1 and 2 week lag in data
6. Validate representative group
    a. Using the normalization method and correlation function compare confirmed influenza data to representative age, borough, and age-and-borough combination subgroup
    b. Extract Season Correlation Data from Results
    c. Identify max correlation value from each subpopulation in each time period (In Sync, 1 Week Lag, 2 Week Lag)
    d. 1 Week lag means comparing weeks 2 to 51 for a subpopulation and weeks 3 to 52 for the total population
    e. Similarly, 2 Week lag means comparing weeks 1 to 50 for a subpopulation and weeks 3 to 52 for the total population
    f. All Correlation values are recorded in one table

g. Process is the same for peaks except the analysis considers only the week with the peak in the total population and the previous five weeks

h. Assume the peak is in week 7

i. 2 Week lag means a subpopulation compares weeks 1,2, 3, 4, & 5 to the Total population's weeks of 3, 4, 5, 6, & 7

j. 1 Week lag means a subpopulation compares weeks 2, 3, 4, 5, & 6 to the Total population's weeks of 3, 4, 5, 6, & 7

k. The total population time period does not change, the subpopulation time period is shift to compare data from different weeks.

l. The same process is completed for borough data

References Cited

1. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011, April). Predicting flu trends using twitter data. In Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on (pp. 702-707). IEEE.

2. Ajelli, M., & Litvinova, M. (2017). Estimating contact patterns relevant to the spread of infectious diseases in Russia. Journal of Theoretical Biology, 419, 1-7.

3. Andersen, R., & Newman, J. F. (2005). Societal and individual determinants of medical care utilization in the United States. The Milbank Quarterly, 83(4), Online-only.

4. Andersen, R. M. (2008). National health surveys and the behavioral model of health services use. Medical care, 46(7), 647-653.

5. Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. "Twitter catches the flu: detecting influenza epidemics using Twitter." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.

6. Azarian, T., Winn, S., Zaheer, S. A., Buehler, J., & Hopkins, R. S. (2009). Utilization of syndromic surveillance with multiple data sources to enhance public health response. Advances in Public Health Surveillance, 7, 1-6.

7. Badham, J., & Stocker, R. (2010). The impact of network clustering and assortativity on epidemic behaviour. Theoretical population biology, 77(1), 71-75.

8. Ball, F., Britton, T., House, T., Isham, V., Mollison, D., Pellis, L., & Tomba, G. S. (2015). Seven challenges for metapopulation models of epidemics, including households models. Epidemics, 10, 63-67.

9. Ball, F., Sirl, D., & Trapman, P. (2010). Analysis of a stochastic SIR epidemic on a random network incorporating household structure. Mathematical Biosciences, 224(2), 53-73.
10. Beckley, R., Weatherspoon, C., Alexander, M., Chandler, M., Johnson, A., & Bhatt, G. S. (2013). Modeling epidemics with differential equation.
11. Beretta, E., & Takeuchi, Y. (1995). Global stability of an SIR epidemic model with time delays. Journal of mathematical biology, 33(3), 250-260.
12. Brownstein, J. S., Chu, S., Marathe, A., Marathe, M. V., Nguyen, A. T., Paolotti, D., ... & Tizzoni, M. (2017). Combining participatory influenza surveillance with modeling and forecasting: Three alternative approaches. *JMIR public health and surveillance*, *3*(4), e83.
13. Centers for Disease Control and Prevention. Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks. 2004. US Department of Health and Human Services: Atlanta Google Scholar.
14. Centers for Disease Control and Prevention. (2018) Flu View. http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html
15. Centers for Disease Control and Prevention. “Information on 2009 H1N1 Impact by Race and Ethnicity.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 24 Feb. 2010, www.cdc.gov/h1n1flu/race_ethnicity_qa.htm.
16. Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature." Geoscientific model development 7.3 (2014): 1247-1250.
17. Coburn, B. J., Wagner, B. G., & Blower, S. (2009). Modeling influenza epidemics

and pandemics: insights into the future of swine flu (H1N1). *BMC medicine*, *7*(1), 30.

18. Cook S, Conrad C, Fowlkes AL, Mohebbi MH (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLoS ONE 6(8): e23610. https://doi.org/10.1371/journal.pone.0023610
19. Cooper, D. L., Verlander, N. Q., Elliot, A. J., Joseph, C. A., & Smith, G. E. (2007). Can syndromic thresholds provide early warning of national influenza outbreaks?. Journal of Public Health, 31(1), 17-25.
20. Cox, N. J., Brammer, T. L., & Regnery, H. L. (1994). Influenza: global surveillance for epidemic and pandemic variants. European journal of epidemiology, 10(4), 467-470. Chicago
21. Del Valle, S. Y., Hyman, J. M., & Chitnis, N. (2013). Mathematical models of contact patterns between age groups for predicting the spread of infectious diseases. Mathematical biosciences and engineering: MBE, 10, 1475.
22. Del Valle, S. Y., Hyman, J. M., Hethcote, H. W., & Eubank, S. G. (2007). Mixing patterns between age groups in social networks. Social Networks, 29(4), 539-554.
23. Donaldson, L. J., Rutter, P. D., Ellis, B. M., Greaves, F. E., Mytton, O. T., Pebody, R. G., & Yardley, I. E. (2009). Mortality from pandemic A/H1N1 2009 influenza in England: public health surveillance study. Bmj, 339, b5213.
24. d’Onofrio, A. (2005). On pulse vaccination strategy in the SIR epidemic model with vertical transmission. Applied Mathematics Letters, 18(7), 729-732.
25. Doroshenko, A., Cooper, D., Smith, G., Gerard, E., Chinemana, F., Verlander, N., & Nicoll, A. (2005). Evaluation of syndromic surveillance based on National Health

Service Direct derived data–England and Wales. MMWR Morb Mortal Wkly Rep, 54(Suppl), 117- 122.

26. Edmunds, W. J., O'callaghan, C. J., & Nokes, D. J. (1997). Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. Proceedings of the Royal Society of London B: Biological Sciences, 264(1384), 949-957.
27. Fleming, D. M., M. Zambon, and A. I. M. Bartelds. "Population estimates of persons presenting to general practitioners with influenza-like illness, 1987–96: a study of the demography of influenza-like illness in sentinel practice networks in England and Wales, and in The Netherlands." Epidemiology & Infection 124.2 (2000): 245-253.
28. Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A., & Merler, S. (2012). Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. PLoS computational biology, 8(9), e1002673.
29. Goeyvaerts, N., Willem, L., Van Kerckhove, K., Vandendijck, Y., Hanquet, G., Beutels, P., & Hens, N. (2015). Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence. Epidemics, 13, 1-9.
30. Hall, I. M., Gani, R., Hughes, H. E., & Leach, S. (2007). Real-time epidemic forecasting for pandemic influenza. Epidemiology & Infection, 135(3), 372-385.
31. Hopkins, R. (2012). Syndromic Surveillance [PowerPoint Slides]. Retrieved from http://www.dhhr.wv.gov/oeps/Documents/2012%20Symposium%20Slides/WV%20Synd romic.pdf.

32. Jefferson, H., Dupuy, B., Chaudet, H., Texier, G., Green, A., Barnish, G., ... & Meynard,

J. B. (2008). Evaluation of a syndromic surveillance for the early detection of outbreaks among military personnel in a tropical country. Journal of Public Health, 30(4), 375-383.

33. Jacobsen, S. J., Guess, H. A., Panser, L., Girman, C. J., Chute, C. G., Oesterling, J. E., & Lieber, M. M. (1993). A Population-Based Study of Health Care^ Seeking Behavior for Treatment of Urinary Symptoms: The Olmsted County Study of Urinary Symptoms and Health Status Among Men. Archives of family Medicine, 2(7), 729.

34. Jia, K., & Mohamed, K. (2015). Evaluating the use of cell phone messaging for community Ebola syndromic surveillance in high risked settings in Southern Sierra. *African health sciences*, 15(3), 797-802.

35. Josseran, L., Fouillet, A., Caillère, N., Brun-Ney, D., Ilef, D., Brucker, G., Medeiros, H. & Astagneau, P. (2010). Assessment of a syndromic surveillance system based on morbidity data: results from the Oscour® Network during a heat wave. PloS one, 5(8), e11984.

36. Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. Journal of the Royal Society Interface, 2(4), 295-307.

37. Kermack, W. O., & McKendrick, A. G. (1927, August). A contribution to the mathematical theory of epidemics. In Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences (Vol. 115, No. 772, pp. 700-721). The Royal Society.

38. Krumkamp, R., Kretzschmar, M., Rudge, J.W., Ahmad, A., Hanvoravongchai, P., Westenhoefer, J., Stein, M., Putthasri, W. and Coker, R. (2011). Health service resource needs for pandemic influenza in developing countries: a linked transmission dynamics, interventions and resource demand model. Epidemiology & Infection, 139(1), 59-67.

39. Kumar, S. (2018, March). The Differences Between Machine Learning And Predictive Analytics. Retrieved https://www.digitalistmag.com/digital-economy/2018/03/15/differences-between-machine-learning-predictive-analytics-05977121

40. Lee, J., Kim, J., & Kwon, H. D. (2013). Optimal control of an influenza model with seasonal forcing and age-dependent transmission rates. Journal of theoretical biology, 317, 310-320.

41. Lee, Y. J., Boden-Albala, B., Larson, E., Wilcox, A., & Bakken, S. (2014). Online health information seeking behaviors of Hispanics in New York City: a community-based cross- sectional study. Journal of medical Internet research, 16(7).

42. Lekone, P. E., & Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. Biometrics, 62(4), 1170- 1177.

43. Lombardo, J. S., & Buckeridge, D. L. Disease surveillance: a public health informatics approach. 2007. *Hoboken: Wiley-Interscience*, 41-262.

44. Lombardo, J. S., Burkom, H., & Pavlin, J. (2004). ESSENCE II and the framework for evaluating syndromic surveillance systems. *Morbidity and Mortality Weekly Report*, 159- 165.

45. May, L., Katz, R. L., Test, E., & Baker, J. (2011). Applications of syndromic surveillance in resource poor settings. *World Medical & Health Policy*, 3(4), 1-29.

46. McCluskey, C. C. (2010). Complete global stability for an SIR epidemic model with delay—distributed or discrete. *Nonlinear Analysis: Real World Applications, 11*(1), 55- 59.

47. Metzger, K. B., Hajat, A., Crawford, M., & Mostashari, F. (2004). How many illnesses does one emergency department visit represent? Using a population-based telephone survey to estimate the syndromic multiplier. Morbidity and Mortality Weekly Report, 106-111.

48. Mostashari, F., & Hartman, J. (2003). Syndromic surveillance: a local perspective. Journal of Urban Health, 80, i1-i7.

49. New York City Department of Health and Mental Hygiene. (2015). Syndromic Surveillance Data 2008-2015: Influenza-like illness (ILI). Retrieved from https://a816- healthpsi.nyc.gov/epiquery/Syndromic/

50. Ogunjimi, B., Hens, N., Goeyvaerts, N., Aerts, M., Van Damme, P., & Beutels, P. (2009). Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. Mathematical biosciences, 218(2), 80-87.

51. Oshitani, H., Kamigaki, T., & Suzuki, A. (2008). Major issues and challenges of influenza pandemic preparedness in developing countries. Emerging infectious diseases, 14(6), 875.

52. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). Dataset shift in machine learning. The MIT Press.

53. Read, J. M., Lessler, J., Riley, S., Wang, S., Tan, L. J., Kwok, K. O., Guan, Y., Jiang, Q. J., & Cummings, D. (2014). Social mixing patterns in rural and urban areas of southern China. Proceedings of the Royal Society of London B: Biological Sciences, 281(1785), 20140268.

54. Ortiz, J.R., Zhou H., Shay D.K., Neuzil K.M., Fowlkes A.L., & Goss C.H. (2011) Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends. PLoS ONE 6(4): e18687. https://doi.org/10.1371/journal.pone.0018687

55. Pineda, A. L., Ye, Y., Visweswaran, S., Cooper, G. F., Wagner, M. M., & Tsui, F. R. (2015). Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. Journal of biomedical informatics, 58, 60-69.

56. Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., & Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS computational biology, 11(10), e1004513.

57. Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R., & Nagler, A. (2014). Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT. *Bone marrow transplantation*, *49*(3), 332-337.

58. Silva, J. C., Shah, S. C., Rumoro, D. P., Bayram, J. D., Hallock, M. M., Gibbs, G. S., & Waddell, M. J. (2013). Comparing the accuracy of syndrome surveillance systems in detecting influenza-like illness: GUARDIAN vs. RODS vs. electronic medical record reports. *Artificial intelligence in medicine*, 59(3), 169-174.

59. Skvortsov, A., Ristic, B., & Woodruff, C. (2010, July). Predicting an epidemic based on syndromic surveillance. In Information Fusion (FUSION), 2010 13th Conference on (pp. 1-8). IEEE.

60. Soto, G., Araujo-Castillo, R. V., Neyra, J., Fernandez, M., Leturia, C., Mundaca, C. C., & Blazes, D. L. (2008, November). Challenges in the implementation of an electronic surveillance system in a resource-limited setting: Alerta, in Peru. *In BMC proceedings*

(Vol. 2, No. Suppl 3, p. S4). BioMed Central Ltd.

61. Stoto, M. A. (2005). Syndromic surveillance. *Issues in Science and Technology*, 21(3), 49-56.

62. Thompson, W. W., Comanor, L., & Shay, D. K. (2006). Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease. The Journal of infectious diseases, 194(Supplement_2), S82-S91.

63. Ukwaja, K. N., Alobu, I., Nweke, C. O., & Onyenwe, E. C. (2013). Health care-seeking behavior, treatment delays and its determinants among pulmonary tuberculosis patients in rural Nigeria: a cross-sectional study. *BMC health services research*, *13*(1), 1.

64. Uscher-Pines, L. O. R. I., Duggan, P. S., Garoon, J. P., Karron, R. A., & Faden, R. R. (2007). Planning for an influenza pandemic: social justice and disadvantaged groups. Hastings Center Report, 37(4), 32-39.

65. Valle D, Clark JS, Zhao K (2011) Enhanced Understanding of Infectious Diseases by Fusing Multiple Datasets: A Case Study on Malaria in the Western Brazilian

Amazon Region. PLoS ONE 6(11): e27462. doi:10.1371/journal.pone.0027462

66. van der Hoeven, M., Kruger, A., & Greeff, M. (2012). Differences in health care seeking behaviour between rural and urban communities in South Africa. International journal for equity in health, 11(1), 31.

67. Venkatarao, E., Patil, R. R., Prasad, D., Anasuya, A., & Samuel, R. (2012). Monitoring data quality in syndromic surveillance: learnings from a resource limited setting. Journal of global infectious diseases, 4(2), 120.

68. Wallinga, J., Teunis, P., & Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. American journal of epidemiology, 164(10), 936-944.

69. Wiens, J., & Shenoy, E. S. (2017). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149-153.

70. Wierman, J. C., & Marchette, D. J. (2004). Modeling computer virus prevalence with a susceptible-infected-susceptible model with reintroduction. Computational statistics & data analysis, 45(1), 3-23.

71. Worby, C. J., Chaves, S. S., Wallinga, J., Lipsitch, M., Finelli, L., & Goldstein, E. (2015). On the relative role of different age groups in influenza epidemics. Epidemics, 13, 10-16.

72. Xu, Q., Gel, Y. R., Ramirez, L. L. R., Nezafati, K., Zhang, Q., & Tsui, K. L. (2017). Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. PloS one, 12(5), e0176690.