Distribution Agreement

Signature:

______________________________ ______________

Carla Kumbale

Date

An Evaluation of Open Source Tools to Estimate the Reproduction Number of the 2009 Influenza A/H1N1 Pandemic in the USA

By

Carla M. Kumbale

Master of Public Health

Epidemiology

Benjamin Lopman, PhD, MSc

Faculty Thesis Advisor

Matthew Biggerstaff, ScD, MPH

Field Thesis Advisor

An Evaluation of Open Source Tools to Estimate the Reproduction Number of the 2009 Influenza A/H1N1 Pandemic in the USA

By

Carla M. Kumbale

B.S., Georgia Institute of Technology, 2015

Thesis Committee: Matthew Biggerstaff, ScD, MPH, Benjamin Lopman, PhD, MSc

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2018

Abstract

An Evaluation of Open Source Tools to Estimate the Reproduction Number of the 2009 Influenza A/H1N1 Pandemic in the USA

By: Carla Kumbale

Introduction: The reproduction number (R), which is a key epidemiological parameter to be estimated during emerging outbreaks such as the 2009 influenza pandemic, is computed using the R0 R-package. In addition, another R-package, EpiEstim, is also utilized to compute the instantaneous reproduction numbers (R(t)). Estimates are compared with the literature in order to further validate the use of these R-packages so that institutions such as the Centers for Disease Control and Prevention can implement these tools in order to rapidly compute these parameters during an emergency.

Methods: In the R0 package we use the Maximum Likelihood (ML), Exponential Growth (EG), Time-Dependent (TD), and Sequential Bayesian (SB) methods to compute four different estimates of R. In the EpiEstim package, we use the Bayesian Statistical technique implemented in this R-package to estimate R(t). A serial interval of 3.6 days with a standard deviation of 1.6 days is assumed for all of the estimates. These reproduction numbers are then compared to the literature.

Results: Several estimates are computed through the use of the R0 package. For the ML method, we estimate R values that vary between 1.41 to 3.54 depending on the selected time period of incidence cases. For the TD, EG, and SB methods, an R value of 1.88, 1.91, and 1.24 are computed respectively. Finally, the R(t) values, computed through the use of the EpiEstim package vary between 1.04-3.37 (weekly time window) and 1.24 – 3.33 (10-day time window). These values are subsequently compared with the results found in the literature.

Conclusion: This study has both demonstrated the use of and further validated the two computer packages that are now available for general use by non-modelers. Although the use of these packages still requires a certain minimum knowledge of statistical methods, the availability of these packages vastly improves the tools now at the disposal of public health practitioners during an epidemic/pandemic.

An Evaluation of Open Source Tools to Estimate the Reproduction Number of the 2009 Influenza A/H1N1 Pandemic in the USA

By

Carla Kumbale

B.S., Georgia Institute of Technology, 2015

Thesis Committee: Matthew Biggerstaff, ScD, MPH, Benjamin Lopman, PhD, MSc

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2018

Acknowledgements

I would like to gratefully thank my thesis advisors Dr. Matthew Biggerstaff and Dr. Ben Lopman for all their support, encouragement, and guidance they provided me. It was truly a pleasure to work with you both.

I would like to sincerely thank Dr. Lance Waller for introducing me to Dr. Biggerstaff and also for being a constant support system and wonderful mentor.

I would like to express my sincere appreciation to Dr. Eberhard Voit and Dr. Qiang Zhang for their continuous support and guidance through all my endeavors. You both are truly an inspiration for me.

Most of all, I thank my parents. I thank my father for always supporting me and for teaching me the meaning of scientific and mathematical rigor by his own example and I thank my mother who has always been my number one supporter.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION

In early April 2009, the public became aware of a novel A/H1N1 influenza virus, also known as “swine flu”. This disease originated from Central Mexico, which soon spread to the United States. [1] In the following weeks, the Director-General of the World Health Organization (WHO) declared this H1N1 influenza outbreak to be an international public health emergency. By July 11th, 2009, the WHO declared a Phase 6 alert, meaning the world was now at the start of the 2009 influenza pandemic. [2] At the conclusion of this pandemic, approximately 60.8 million cases, 274,304 hospitalizations, and 12,469 deaths occurred within the United States. [3]

To more effectively mitigate against the effects of an emerging pandemic such as this, estimation of key epidemiological parameters, such as the serial interval and the reproduction number is critical. [4]

The serial interval is defined as the time between successive cases in a chain of transmission. [5] This is essential for outbreak studies since investigators are able to identify epidemiological links between cases and also to diagnose new cases that have links with laboratory- confirmed cases. [6] In this thesis, a serial interval of 3.6 with a standard deviation of 1.6 is used which was previously estimated by Cowling et al. [7]

The basic reproduction number R_0 is the expected number of secondary cases produced by a single typical infectious case. [8] Furthermore, this parameter describes the average amount of persons a case will infect assuming the entire population is susceptible to infection. [9]

When R_0 is greater than 1, the infection has the propensity to spread across a population. As R_0 increases, the harder it is to control the disease. In contrast, when R_0 is less than 1, this usually indicates that there will be a decline of infection in the population however, exceptions exist. [9] R_0 is mathematically defined as follows: [10]

$$R_0 = \tau * \beta * \delta$$

Where:

τ = the transmission probability of the disease between susceptible and infected individuals

β = the rate of contact between susceptible and infected individuals per unit time

δ = the duration and infected person remains infectious to other susceptible individuals

The instantaneous reproduction number R(t) also termed the effective reproduction number, as defined by Fraser et al., is the number of secondary cases on average that each infected individual would infect if conditions remained as they were at each time step denoted as *t*. [11] This parameter is calculated in a population with underlying immunity therefore, accounting for reduced susceptibility to infection within a population. [12] R(t) is mathematically defined as the ratio of the number of new infections occurring at time step *t* to the sum of infectious incidence ending on time step t-1 weighted on the infectivity function which can be seen below: [13]

$$\mathrm{R(t)} = \frac{I_t}{\sum_{s=1}^{t} I_{t-s}\, W_s}$$

Where:

I_t = the number of infections generated at time step *t.*

W_s = the infectivity function

Several methods exist to estimate infectious disease transmission parameters. [14] During the 2009 pandemic, over 78 separate estimates for R were made. [9] Because no open-source tools were available at that time that could be executed quickly and easily by public health officials with limited modeling experience, these estimates often represented a substantial investment of time and expertise by mathematical modelers. [14] For public health institutions such as the Centers for Disease Control and Prevention (CDC), which play an active role during a pandemic, accurate, reproducible, and fast estimates of parameters such as the reproduction number and serial interval are of crucial importance.

Recently, two open-source R-program packages have been developed that can be used to estimate the reproduction number and instantaneous reproduction numbers during influenza and other infectious disease outbreaks: R0 and EpiEstim. [13,14,15] Currently, these two packages are the only open-source R=packages available that can be used to estimate these parameters to our knowledge. EpiEstim supplies a framework which can estimate the instantaneous reproduction number from the traditional epidemic curve while the package R0 can implement five different methods to estimate the reproduction number based on incidence data. These methods include: Attack Rate (AR), Exponential Growth (EG), Maximum Likelihood (ML), Time-Dependent (TD), and Sequential Bayesian (SB). [14] The motivation behind the creation of these packages was to provide public health practitioners with a ready-to- use tool that only requires data that are commonly collected or estimated during an outbreak, such as the epidemic curve and serial interval. [13,14] The packages potentially would allow for the rapid computation of key parameters during an emerging pandemic by CDC and other public health agency personnel, strengthening the evidence base for selection and use of pharmaceutical and non-pharmaceutical interventions by these organizations.

However, while the R0 package has been validated using incidence data from the 1918 pandemic, a validation of this packages comparing results against each of the different methods used to estimate the reproduction number and against the results produced by the White et al. (2009) paper, which used the same 2009 pandemic line list data used in this analysis, and a meta- analysis of pandemic 2009 reproduction numbers conducted by Biggerstaff et al. (2014) has not yet been done. [9,16] White et al. use a likelihood- based method to estimate the reproduction number. Data consisted of date of illness onset and date of report to the CDC from early reported influenza cases in the United States during the 2009 pandemic. [16] When estimating the R value alone, the authors made use of previous estimates of the mean of the serial interval provided by Cowling at al. (3.6 days) and Fraser et al. (1.91 days). [7,17] Using these estimates of the serial interval, White et al. calculate R which ranges from 1.5 to 3.1.[16] In the systematic literature review by Biggerstaff et al., the summarized R value estimate of the 2009 pandemic was found to be 1.46 (IQR: 1.30–1.70), which is the median point estimate in the community setting for all waves of this illness computed from 78 separate estimates of the reproduction number. [9]

The validation of this package against estimates produced from the same dataset and compared to the consensus pandemic reproduction number would improve public health officials' confidence in the use of these packages during the early stages of a pandemic.

Finally, to our knowledge, R(t) has not yet been estimated for the 2009 influenza pandemic for the entire population of the USA. However, two studies mentioned in the Biggerstaff et al. meta- analysis estimate this parameter within a camp based setting (R(t): 1.4-3.3) and a school setting (R(t)= 1.3-2.0) in the USA. [9,18,19] Therefore, we aim to estimate the R(t) values for this pandemic encompassing the entire USA by utilizing the EpiEstim R-package in this thesis.

Purpose

The work in this thesis is aimed at demonstrating the use of the *R0* package to estimate the reproduction number during an emerging influenza pandemic. Results obtained are compared against reproduction number estimates from the White et al. (2009) and the Biggerstaff et al (2014) papers to further validate this tool. [9,16] In addition, we compare the results produced by the different techniques utilized by the R0 package which include the following: ML, SB, TD, and EG. Finally, we utilize the R-package EpiEstim to estimate the instantaneous reproduction numbers for this 2009 Influenza pandemic in the USA. [13,14]

Specific Aims:

The three specific aims of this thesis are to:

1) **Aim 1:** Estimate the reproduction number R using a maximum likelihood estimator (ML) implemented in the R0 package and compare these results with the results obtained in the literature. [16]
2) **Aim 2:** Compute a reproduction number R for a single study period starting from the earliest beginning date cited in White et. al. (March/28/2009) and ending on the latest date (May/3/2009). Each of the four applicable methods implemented in the R0 package which are valid for this particular investigation: EG, TD, ML, and SB are used for this computation. We then aim to make comparisons of these results with the results found in the literature.
3) **Aim 3:** Estimate the instantaneous reproduction number R(t) for this pandemic in the USA through the use of a Bayesian statistical method included in the EpiEstim package.

CHAPTER TWO: METHODS

Data

We use an updated version of the data used in the White et al. paper provided by the CDC which includes a line list of reported cases of influenza A/H1N1 in the United States. Information on approximately 1880 confirmed and probable cases were used. Of these, 1194 confirmed and probable cases had both a date of onset with a date of report to the CDC. The distribution of confirmed and probable cases can be seen in Figure 1. We include probable cases in this analysis since greater than 90% of these cases were later confirmed. [16] Individual-based data becomes considerably less frequent in favor of aggregate counts of new cases after May 13th. [16] Therefore, we chose the range of onset dates included in this analysis to be March 28th, 2009 to May 3, 2009 and the range of report dates to be April 24, 2009 to May 8, 2009.

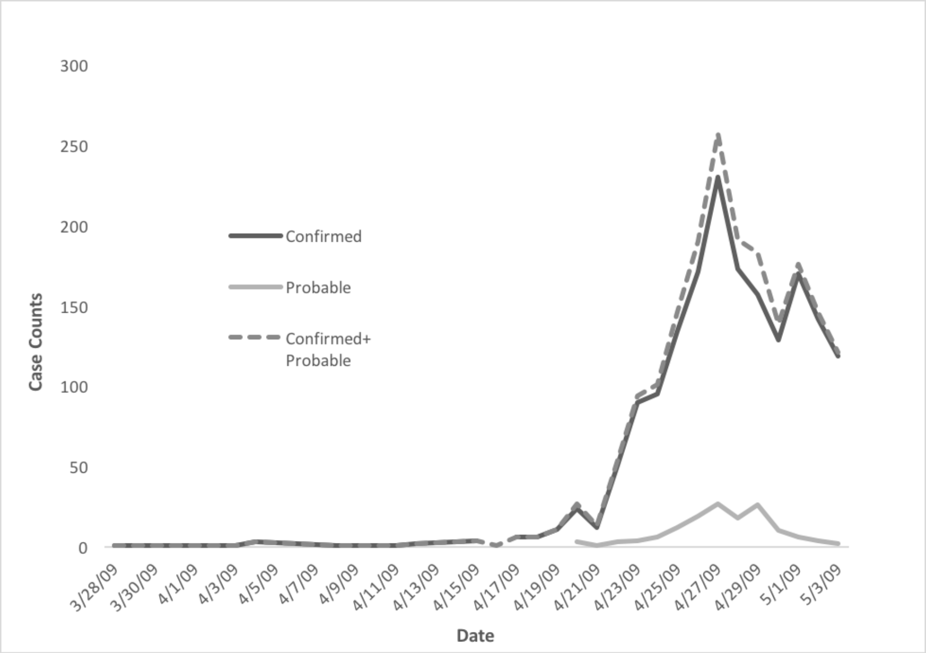

Figure 1: Confirmed and probable cases plotted by onset time occurring in the United States. First date of onset is March 28, 2009 ending on May 3, 2009. It can be observed that onset times rapidly decline as we approach the final onset date. There were 1738 confirmed cases, 142 probable cases, and a total of 1880 confirmed and probable cases.

The analysis is performed using imputed data where if the date of onset is missing, it is imputed from the date of report. Shown in Figure 2 is the average time between onset of symptoms and case report to the CDC. If a report date is known but the onset date is missing, we compute an estimated date of onset through the following equation:

$$\omega = R - A$$

Where:

ω = The estimated onset date

R = The report date

A= The average time in days between onset and date of report to the CDC

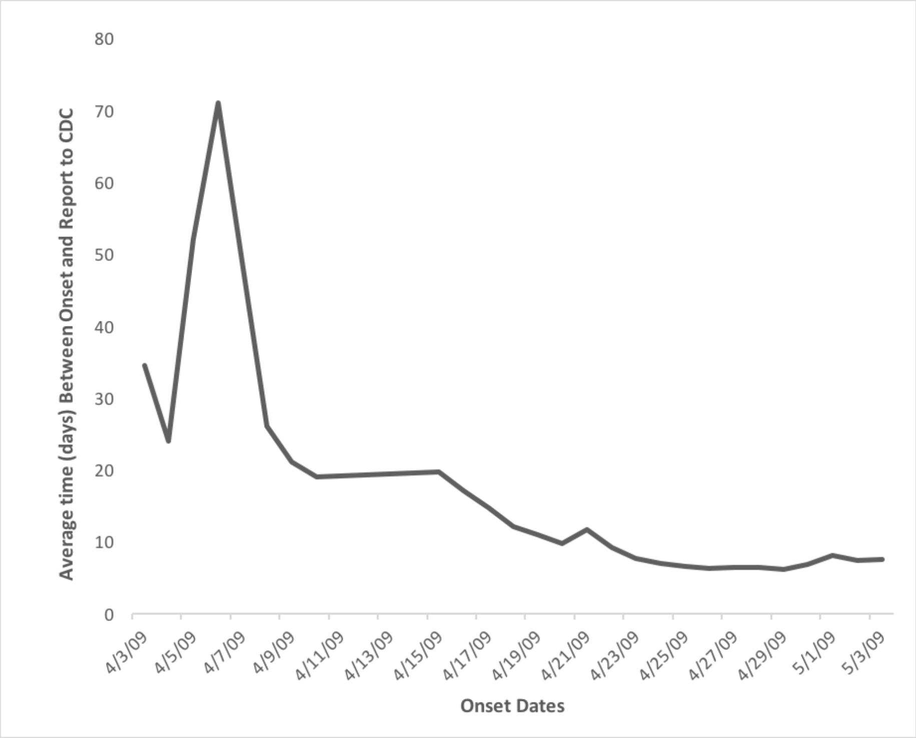

Figure 2: The average reporting delay by days to the CDC of cases with a known date of onset is shown. It can be observed that the peak reporting delay between onset and report date occurs on April 6, 2009. As time progresses, reporting delay rapidly declines.

R0 Package

Estimates of the reproduction number are computed through the use of the R-package R0. We stratify the results for four datasets: data with an onset date on or before 4/25/09,4/26/09, 4/27/09, or 4/28/09 as done by the White et al. paper. [16] For the stratified data, estimates are made through the use of the maximum likelihood estimation model proposed by White & Pagano. [20] This method is also used to estimate the R value in the White et al. paper as well. [16]

Estimates of the overall R value which consists of data with an onset date for the overall time period are also computed through the use of the R-package R0. The range of the overall time period for the dataset is from 3/ 28/2009 to 5/3/2009. We implement four different methods which include: ML, EG, SB, and, TD to compute four estimates. We assume a serial interval of 3.6 days with standard deviation of 1.6 days for all estimates. [7]

EpiEstim Package

Estimates of the instantaneous reproduction number, R(t), were computed through the use of the R-package EpiEstim. Estimates are made for an epidemic over predefined time windows through the use of a Bayesian statistical method proposed by Wallinga & Teunis. [21] We define a 7-day time window and 10-day time window for which estimates are made. The first week begins on April 18, 2009 and ends on April 25, 2009. The last week defined begins on April 26, 2009 ending on May 3, 2009. Additionally, the first 10-day period begins on April 15, 2009 and ends on April 25, 2009. The last 10-

day period begins on April 23, 2009 and end on May 3, 2009. Again, for all analyses, we assume a mean serial interval of 3.6 with a standard deviation of 1.6. [7]

Description of the Mathematical Models

The two computer packages, R0 and EpiEstim utilized in this thesis take somewhat distinct approaches in order to compute the reproduction number. The first package R0 computes a R value for the entire study period whereas the second computes R(t) over a time window τ but with reproduction number considered constant over the specified time window.[14] This package additionally makes available several models for computing R.[14] These methods are termed Maximum Likelihood Estimator (ML), Exponential Growth (EG), Sequential Bayesian (SB) and Time Dependent (TD) reproduction number.[14] The EpiEstim models the reproduction number as a RV (Random Variable), which is also done in SB in R0 package but with the caveat that in the former, R(t) is considered constant over a time window τ and uses a Bayesian statistical method from a Gamma Prior. [13]

Each of these models are described below:

R0 - Maximum Likelihood Estimator (ML):

This model developed by White and Pagano assumes that each primary case generates secondary cases whose expected value is the reproduction number. [20] Furthermore, it also assumes that the primary cases on the first day, generate the subsequent secondary cases over the assumed maximal serial interval. For each of the days in the study period in turn generate secondary cases over the following days in the serial interval. Given the number of incidence cases during the study period, this method computes the reproduction number that will maximize the likelihood function.

This analysis assumes that the epidemic curve is analyzed from the very first day. If that is not the case, the initial R value will be over estimated. This model is built upon the fundamental assumption that the number of secondary cases produced by each infected individual is Poisson distributed. [20] In addition, it also assumes that this whole process, which starts from a seed number of cases is a closed system. The assumption of the process being a closed system may not be totally realistic if the disease is dispersed across the globe. The results from this method are more reliable when the knowledge of the serial interval is well known. [20]

For simplicity sake, t implicitly refers to one day in this thesis. Thus, if N_t, t=1, 2, 3, …T represents the set of incidence cases generated over a period with the maximal length of the serial interval being k, then N_0 are initial seed cases that initiate the whole process of generating subsequent incidence of cases N_t. [20]

Since production of secondary cases from a single primary case is assumed to be Poisson distributed with parameter R, it then follows that the RV X_t representing the total number of secondary cases generated over the maximal serial interval at time t, are also Poisson distributed with parameter $\lambda_t=N_tR$. [20]

Additionally, it is assumed that the distribution of cases X_t over the forward serial interval denoted by $X_{ts,}$ for t=0,1,2,3…T and s=1,2…k are themselves RVs of a multinomial distribution with unknown probability p. That is X_{ts}, ~ Mult (k, p). Thus $X_t = \sum_{j=1}^{k} X_{tj.}$ Moreover, X_{ts} are secondary cases generated on day s from N_t on day t where t=0,1, 2,….,T and s=1,2,…,k where k is the maximal length of the serial interval. [20]

In this this study, based upon documented data from the CDC, the maximum serial interval, k, is known. However, k is varied depending upon the length of the incidence data vector used for a particular study.

The overall likelihood is a product of individual likelihoods for each t and assuming that the secondary cases generated for each t are independent, likelihood function L takes the following form: [20]

$$L\left(R, X\middle|, N_0, N_1, \dots N_T, p_1, p_{2,} \dots, p_k\right) = \prod_{t=0}^{T}\left(\frac{e^{-\lambda}{}_t \cdot \lambda_t^{X_t}}{X_t!}\right)\left[\left(\frac{X_t!}{X_{t1}!.X_{t2}! \dots X_{tk}!}\right)\prod_{j=1}^{k} p_j^{X_{tj}}\right]$$

This is a product of two groups of terms with first group from the Poisson PDF and a second bracketed term from the multinomial distribution over the serial interval k. The p terms of a multinomial distribution are constrained by the requirements imposed by the first and second axioms of probability.

There are further simplifications that can be made by observing the final k terms are unobservable. Hence the final k terms can be dropped which reduces the number of terms to

$T - k$ when T>>k. Thus, the future terms of X_{ts} for t>T and S>0 are unknown and hence the last several X_t terms are normalized to conform to the requirement of multinomial or binomial RVs.

By making substitutions, the likelihood reduces to the following compact equation: [20]

$$\mathrm{L}(\mathrm{R}|\mathbf{p}) = \prod_{t=1}^{T} \frac{e^{-\propto_t} \propto_t^{N_t}}{N_t!} \qquad (1)$$

$$where \ \propto_t = R \sum_{i=1}^{\min(k,t)} N_{t-i}\, p_i$$

Since logarithmic functions are monotonic, the original function can be maximized by maximizing the logarithm of the likelihood function. This has the advantage of transforming the multiplicative terms to one of summation. Taking the natural logarithm of (1) and simplifying, the likelihood function becomes: [20]

$$\log \mathrm{L}(R_0|\mathbf{p}) = l(R_0|\mathbf{p}) = -\propto_t + \sum_{t=1}^{T} N_t \log \propto_t - constant$$

This function is a maximum for the unknown parameter R$_0$ when $\frac{dl}{dR} = 0$

$$\frac{dl}{dR} = -\sum_{t=1}^{T} \sum_{i=1}^{\min(k,t)} N_{t-i}\, p_i + \frac{\sum_{t=1}^{T} N_t}{R_0} = 0$$

$$\hat{R} = \frac{\sum_{t=1}^{T} N_t}{\sum_{t=1}^{T} \sum_{i=1}^{\min(k,t)} N_{t-i}\, p_i}$$

R0 - Exponential Growth Method (EG)

This method developed by Wallinga and Lipstich, uses a Euler-Lotka equation in demography which is based on the population growth of females, since the female population is considered as the limiting factor. [22] Here the growth rate of females in the Euler-Lotka equation is interpreted as the growth in infection. The serial interval distribution is assumed known and using this information R is computed for the whole study interval. The EG method requires a scrutiny of the epidemic curve and requires the identification of the period in the epidemic curve where the growth is exponential. No assumption is made on mixing in the population. These equations presented are defined: [22]

$$b(t) = \int_0^\infty e^{-ta} b(t) n(a) da \qquad (3)$$

Where b(t) = population birth rate at time t

n(a) = rate of production of female-offspring

by dividing equation (3) by b(t) which can be taken out of the integral then the equation reduces to: [22]

$$1 = \int_0^\infty e^{-ta} n(a) da \qquad (4)$$

If n(a) is integrated over the whole lifespan we obtain the total number of female offspring produced by a single mother over her lifespan.: [22]

$$R = \int_0^\infty n(a) da$$

Although the upper limit on the integral in reality is finite, it can be replaced by ∞ without any loss of generality. Additionally, R can now be interpreted as the reproduction number.

If n(a) is now normalized as below:[22]

$$g(a) = \frac{n(a)}{\int_0^\infty n(a) da} = \frac{n(a)}{R}$$

If we now interpret the ‘age’ of an infection to be the time since infection, then the generation interval distribution is equivalent to g(a) above. Substituting the equation above in (4) we obtain: [22]

$$\frac{1}{R} = \int_{0}^{\infty} e^{-ta} g(a) da$$

The right hand side (RHS) of the equation above can be readily recognized as the one-sided Laplace Transform of the function g(a). At the same time, the moment generating function of an RV x is: [22]

$$M(x) = \int_{0}^{\infty} e^{xa} g(a) da$$

Which is the same as the Laplace transform except with a negative parameter -x.

R can then be obtained by using a Moment Generating Function M(-r) where r is the growth rate. [22]

$$R = \frac{1}{M(-r)}$$

Depending upon the assumed probability distribution for generation time which in turn is based upon the disease under study, then different moment generating functions result. More details are available in the paper by Wallinga and Lipsitch. [22]

R0 - Sequential Bayesian Method (SB)

The Bayesian statistical method proposed by Bettencourt and Ribeiro, is used estimate R. [23] The infectiousness at a future time which can be the next day is assumed Poisson distributed with a certain mean. The expression for the mean has the term R that is to be estimated and the number of incidences at the present time. Starting from a gamma distributed prior for R the Poisson posterior distribution is updated by applying Bayes theorem as each new incident data is included. This updated posterior becomes the prior distribution for the next update of the posterior for the following day. As more and more incidence data is taken into account R tends to decrease.

This model implicitly uses an exponential distribution for the generation time and in addition it also assumes that there is random mixing in the population. This method will fail when there are gaps in the incidence data, that is periods during which the number of observations is 0. Because of the continuous update made to the prior distribution, it can take into account the impact of intervention on a real-time basis. [23]

This method is similar to the method described for the EpiEstim package (shown below) except for the fact that the posterior distribution updates and becomes the prior distribution for the computation of the reproduction number for following day.[13, 23]

The analysis begins with a non-informative prior on the conditional distribution of R. The conditional distribution of R given a set of number of incidences up to time t, the prior distribution R used on each new day is the posterior distribution from the previous day [23]

$$P(R|\mathrm{N}_0, \mathrm{N}_1, \ldots., \mathrm{N}_t) = \frac{P(\mathrm{N}_t|\mathrm{R}, \mathrm{N}_0, \mathrm{N}_1, \ldots., \mathrm{N}_{t-1})P(R|\mathrm{N}_0, \mathrm{N}_1, \ldots., \mathrm{N}_{t-1})}{P(\mathrm{N}_0, \mathrm{N}_1, \ldots., \mathrm{N}_t)} \ldots\ldots\ldots\ldots\ldots\ldots. (2)$$

More details are available on the mathematical derivation in the Bettencourt and Ribeiro paper. [23]

R0 – Time Dependent Method (TD)

In this method, proposed by Wallinga & Teunis, the probability that a case on day i was infected by a case on prior days j is computed from an assumed serial interval distribution. [21] The reproduction number for a particular day's incidence is then computed as the sum of all probabilities from prior incidences that may contributed to the day's incidence under consideration. An average value of reproduction number is then computed over all cases that had the same date of onset. This method like others in the R0 package assumes that the serial interval distribution is known or at least that it is well understood. The serial interval distribution is dependent upon the outbreak under study. Bias can tend to increase when data is aggregated over long intervals. [21]

This method exploits the fact that serial interval distribution is known for several diseases. Using pairs of cases rather than the entire infection network, it obtains a likelihood-based estimates of reproduction number. Let p_{ij} be the probability that case *i* has been infected by case *j* and let *w*(t) be the known sereial interval distribution, then, given their difference in time of symptom onset $t_i – t_j$, can be expressed in terms of the probability distribution for the generation interval. This distribution for the generation interval is available for many infectious diseases, and we denote it by *w*(τ). The probability that case *i* has been infected by case *j* is then the probability that case *i* has been infected by case *j* divided by the sum of probabilities that case i has been infected by any other case k. Since the distribution of the serial interval is assumed known then the expression for p_{ij}. [21]

$$P_{ij} = \frac{N_i w(t_i - t_j)}{\sum_{i \neq k} N_i w(t_i - t_k)}$$

Where N_i is the number of cases on day i.

Then the reproduction number R_i is then expressed as: [21]

$$R_i = \sum_{\forall j} P_{ij}$$

The average reproduction number can now be calculated for the whole period of study: [21]

$$R = \sum_{\forall i} \frac{R_i}{N_i}$$

The details of the mathematics underlying this whole derivation is provided in the Walling and Teunis paper [21].

EpiEstim - Bayesian Statistical Inference

This package also uses a Bayesian statistical technique, which was developed by Anne Cori et al., except that the reproduction number is assumed to be constant over a subset of period of study termed time window. [13] Thus, this method involves two probability distribution, one for prior and the other for posterior. The choice of time window is important so that it neither too small nor too large. The authors provide a guide for the proper choice of time window where the total number of incidences in that period is greater than a minimum threshold. This model assumes that the transmissibility of the disease over a time window is a Poisson distributed. Infectiousness is assumed to coincide with symptom onset. Since the posterior distribution has an analytical expression, it is possible to link the expression for the posterior CV (standard deviation divided by mean) to the number of incident cases in a time window and then imposing a posterior CV smaller than a required value-threshold. The assumption of infectiousness starts with symptom onset although reasonable for diseases such as influenza, is positively not true for diseases such as HIV.

This method estimates the instantaneous reproduction number $R_{t\tau}$ over a period τ as a Poisson process. [13] The rate at which new secondary infections are generated at time t from a person infected at a prior time $t - s$ is $R_t p_s$ where p_s is the probability of infectiousness at time s. If the time period $t - s$ is characterized by a number of discrete incidences N_s ,s=0,1,…,$t - 1$ then the incidence I_t is Poisson distributed with a mean $R_t \sum_{s=1}^{t} N_{t-s} p_s$.

Let $\lambda_t = \sum_{s=1}^{t} N_{t-s} p_s$, then the likelihood function of the number of incidences being N_t given prior vector of incidences $\mathbf{N_s} = (N_0, N_1, \dots., N_{t-2}, N_{t-1})$: [13]

$$P(N_t|\boldsymbol{N_s}, \boldsymbol{p}, R_t) = \frac{(R_t \lambda_t)^{N_t} e^{-R_t \lambda_t}}{N_t!}$$

Now, for the vector of incidences $\boldsymbol{N_\tau} = (N_{t-\tau+1}, N_{t-\tau+2}, \dots, N_{t-1}, N_t)$ over a span of time τ leading up to time t and assuming a constant reproduction number $R_{t,\tau}$ over that time-span, the likelihood function becomes: [13]

$$P\left(\boldsymbol{N_\tau} \middle| \boldsymbol{N_{t-\tau}}, \boldsymbol{p}, R_{t,\tau}\right) = \prod_{s=t-\tau+1}^{t} \frac{(R_{t,\tau}\lambda_t)^{N_t} e^{-R_{t,\tau}\lambda_s}}{N_s!} \dots\dots\dots\dots\dots\dots .(2)$$

conditional upon a vector of incidences $\boldsymbol{N_{t-\tau}} = (N_0, N_1, \dots, N_{t-\tau})$

Assuming a Prior for $R_{t,\,\tau}$ as gamma distributed with parameters α and β, the prior PDF $P(R_{t,\,\tau})$ is:

$$P(R_{t,\tau}) = \frac{R_{t,\tau}^{\alpha-1} e^{-\frac{R_{t,\tau}}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$$

The posterior probability with this conjugate Prior and with the likelihood function given by (2) is: [13]

$$P\left(\boldsymbol{N_{\tau,}}, R_{t,\tau} \middle| \boldsymbol{N_{t-\tau}}, \boldsymbol{p}\right) = \prod_{s=t-\tau+1}^{t} \frac{(R_{t,\tau}\lambda_t)^{N_t} e^{-R_{t,\tau}\lambda_s}}{N_s!} . \frac{R_{t,\tau}^{\alpha-1} e^{-\frac{R_{t,\tau}}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$$

$$= R_{t,\tau}^{\alpha+\sum_{s=t-\tau+1}^{t}(N_{t-s}-1)} . e^{-R_{t,\tau}[\sum_{s=t-\tau+1}^{t}(\lambda_t+\beta^{-1}])} \prod_{s=t-\tau+1}^{t} \frac{{\lambda_s}^{N_s}}{N_s!} [\Gamma(\alpha)\beta^{\alpha}]^{-1}$$

It can be easily recognized that the posterior PDF is also gamma distributed with parameters: [13]

$$\left(\alpha + \sum_{=t-\tau+1}^{t} N_s\right), \left(\frac{1}{\beta^{-1} + \sum_{s=t-\tau+1}^{t} \lambda_t}\right)$$

The expression for the posterior mean of instantaneous reproduction number over the time-span τ is: [13]

$$R_{t,\tau} = \frac{\alpha + \sum_{=t-\tau+1}^{t} N_s}{\beta^{-1} + \sum_{s=t-\tau+1}^{t} \lambda_t}$$

With this method, the authors take the uncertainty of the serial interval distribution by estimating the instantaneous reproduction number by drawing a specified number of samples, n1, from specified truncated normal distribution around a specified mean. These draws are discretized over the serial interval leading up to a day t. For each sample drawn from the truncated normal, the posterior mean over a time-span is computed. The number of samples to be drawn over the study time-span, n2, can also be specified in the computer package. Out of the total $n1 * n2$

computations of $R_{t,\ \tau}$, a mean of the instantaneous $R_{t,\tau}$ is computed. More details are available in the reference. [13]

Parameter Inputs: *R0* Package

It is shown in Table 1 the parameter inputs that were made for the following methods: ML, EG, SB, and, TD. In addition, we estimate the R value by previously calculated mean and standard deviation serial interval estimates. [16] The serial interval estimate is provided by Cowling *et al.* which has been obtained using household influenza transmission data. [7] Assuming a Weibull model, they estimate the mean serial interval to be 3.6 days with standard deviation of 1.6 days. [7]

Table 1

<table>
<tr><th colspan="3"><i>R0</i> R package Details</th></tr>
<tr><td colspan="2"><b>Methods</b></td><td>est.R0.ML
est.R0.EG
est.R0.SB
est.R0.TD</td></tr>
<tr><th>Parameters</th><th>Definitions</th><th>Inputs</th></tr>
<tr><td>epid</td><td>The epidemic curve</td><td>our dataset which we call casecounts</td></tr>
<tr><td>mGT</td><td>Generation time
or
serial interval</td><td>Mean serial interval is set at 3.6 with standard deviation 1.6
We assume a Weibull distribution</td></tr>
<tr><td>Begin</td><td>Time estimation begins</td><td>We set this at ‘2009-3-28’</td></tr>
<tr><td>End</td><td>Time estimation ends</td><td>We set this at either: [16]
'2009-4-25'<br>'2009-4-26'<br>'2009-4-27’
or
'2009-4-28'</td></tr>
</table>

Included are the various parameters supplied by the R package, R0. We provide the definitions and our inputs that produce the results presented in this thesis. [14]

Parameter Inputs: EpiEstim Package

Shown in Table 2 are the descriptions of the various parameters supplied by the EpiEstim R package, definitions of each of the parameters, and our inputs. For all analyses, we assume a mean serial interval of 3.6 with a standard deviation of 1.6. [7]

Table 2

EpiEstim R package Details		
Parameters	**Definitions**	**Input**
I	Vector of incidence cases	cases
T. Start, T.End	Vector of positive integers consisting of the start and end times of each time-window for with the reproduction number is estimated.	Varied (Refer to appendix for code)
Method	Method in which the serial interval distribution is specified	"UncertainSI"

n1, n2	n1- size of sample of pairs. Must be a positive integer
	n2- size of the sample drawn from each posterior distribution. Must be a positive integer.
		We set this to be:
		n1= 1000
		n2= 1000

Mean.SI	Mean serial interval	3.6
Std.SI	Standard deviation of the serial interval	1.6
Std.Mean	Standard deviation of the distribution from which the serial interval is drawn	1
Min.Mean.SI	Lower bound of serial interval distribution	3.5
Max.Mean.SI	Upper bound of serial interval distribution	3.7
Min.Std	Lower bound of standard deviation of the SI distribution	1.5
Max.Std	Upper bound of standard deviation of the SI distribution	1.7

Included are the various parameters supplied by the R package, EpiEstim. We provide the definitions and our inputs that produce the results presented in this thesis. [13]

CHAPTER 3: RESULTS

We report an epidemic curve by date of onset shown in **Figure 3**. There were 1880 confirmed or probable onset cases with a date of report to the CDC. The first date of onset is March 28th, 2009 with an end date of May 3, 2009. The first date of report is April 24, 2009 ending on May 8, 2009. Overall, there are 29 days of data used in this analysis.

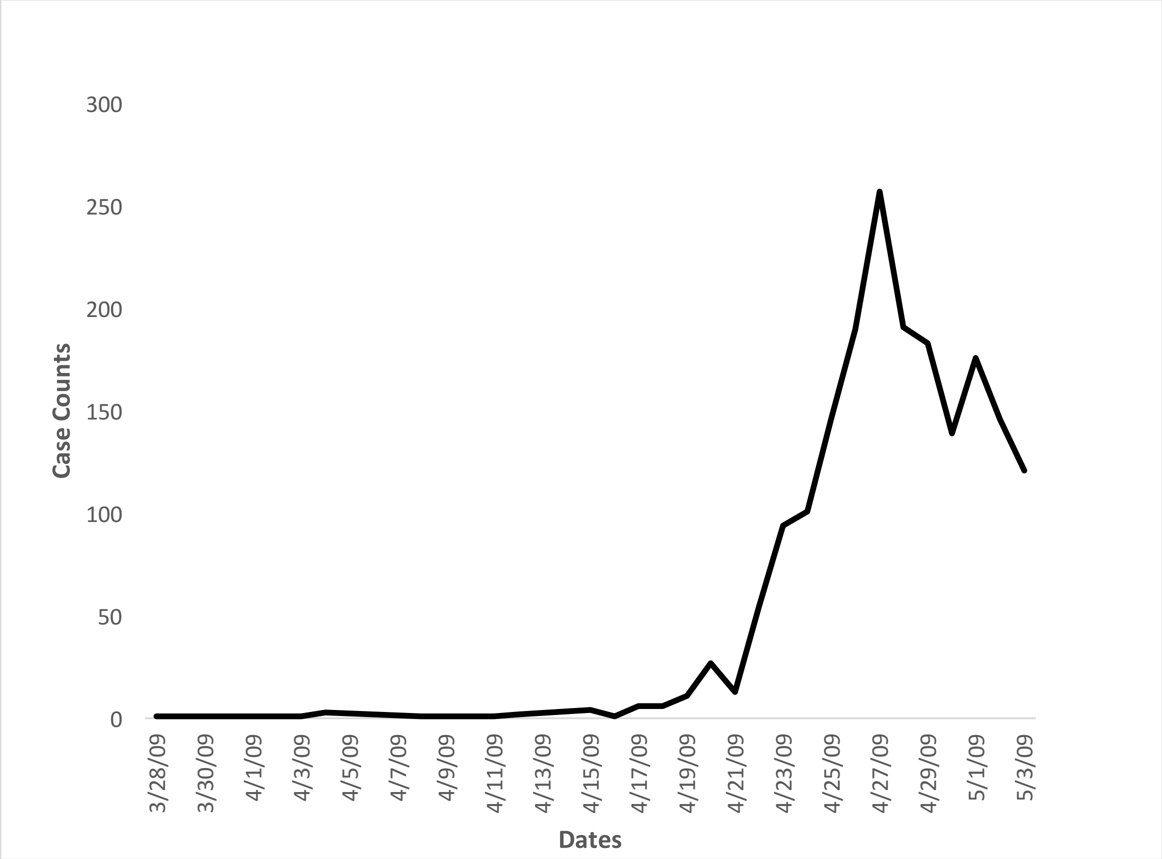

Figure 3. Epidemic curve of infection for 2009 Influenza Pandemic in the United States. The analysis begins on March 28,2009 and ends on May 3, 2009. There are a total of 1880 confirmed and probable cases with a total of 29 days of data used.

As observed in Figure 3, it can be seen that the peak of cases occurred on April 27, 2009. In addition, it can be observed that onset case counts decline rapidly as we approach May

3, 2009, the final date analyzed. This rapid decline of cases is attributable to reporting lag as previously observed by White et al. [16]

Aim 1*: Estimate R_0 using a maximum likelihood estimator (ML) included in the R0 package*

It is shown in Table 3 the R values for the datasets created. Table 3 also includes the results computed by White et al. (2009). [16]

Table 3

Time Period	Mean R_0	2.5th Percentile	97.5^{th} Percentile	Total Cases
Reproduction Numbers mean SI = 3.6 days; σ =1.6				
3/28/09 - 4/25/09	3.54	3.11	4.01	477
3/28/09 - 4/26/09	3.27	2.93	3.64	667
3/28/09 - 4/27/09	3.07	2.80	3.36	924
3/28/09 - 4/28/09	2.58	2.37	2.80	1115
White et al. (2009) Estimates [16]				
Original Data-Reproduction Numbers R_0 mean SI = 3.6 days; σ =1.6				
3/28/09 - 4/25/09	3.48	2.88	3.72	275
3/28/09 - 4/26/09	3.29	2.85	3.47	392
3/28/09 - 4/27/09	2.87	2.55	3.06	529
4/3/09 - 4/28/09	2.59	2.31	2.77	681

R estimates assume the mean of the serial interval (SI) is 3.6 days with standard deviation (σ) of 1.6 days. [7] Estimates computed by White et al. are included. Total case counts for each time period can also be seen.

It can be observed that with a serial interval assumed at 3.6 days, the reproduction numbers vary between 2.58 to 3.54. In addition, as the reporting time period increases, this leads to a decrease in the estimate of the reproduction number. The code implemented in order to achieve these results can be found in the appendix.

Aim 2: *Estimate the overall R_0 using the EG, TD, ML, and SB methods included in the R0 package*

Shown in Table 4 are the R values estimated by the R0 computer package. Four separate values are reported for each of the four methods implements which include ML, EG, SB, and, TD. The range of the overall time period for the dataset is from March 28, 2009 to May 3, 2009.

Table 4

Method	**Reproduction Number** **3/28/2009 – 4/3/2009**	**95% CI**
ML	1.41	[1.32;1.50]
EG	1.88	[1.81;1.95]
TD	1.91	[1.52; 2.31]
SB	1.24	[1.08- 1.40]

Overall R estimates assume the mean of the serial interval (SI) is 3.6 days with standard deviation (σ) of 1.6 days. [7] We implement four methods provided by the *R0* package: ML, EG, TD, and, SB. [14]

The R value reported, in which the TD method is implemented, is an average of the reproduction numbers computed over the 29 days. [14] For the SB method, the reported estimate was obtained on day 29. Reproduction numbers reported all take on a value below 2. The code implemented in order to achieve these results can be found in the appendix.

Aim 3*: Estimate the effective reproduction number R(t) through the use of a Bayesian statistical method included in the EpiEstim package*

Estimates of the instantaneous reproduction number, R(t) are shown in Table 5 in which we define a 7-day time window. The R(t) estimates range from 1.04 to 3.37. As each week progresses in time, both R(t) and the standard deviation of R(t) decreases as expected.

Table 5

Time Period	Mean R_t	SD R_t	2.5th Percentile	97.5th Percentile
Instantaneous Reproduction Numbers R_t **mean SI = 3.6; SD SI = 1.6**				
4/18/09 - 4/25/09	3.37	0.16	3.07	3.69
4/19/09 - 4/26/09	3.12	0.13	2.87	3.37
4/20/09 - 4/27/09	2.55	0.09	2.37	2.74
4/21/09 - 4/28/09	2.12	0.07	1.98	2.26
4/22/09 - 4/29/09	1.73	0.05	1.63	1.84
4/23/09 – 4/30/09	1.49	0.04	1.41	1.58
4/24/09 – 5/1/09	1.32	0.04	1.25	1.39
4/25/09 – 5/2/09	1.17	0.03	1.10	1.23
4/26/09 – 5/3/09	1.04	0.03	0.98	1.09

Instantaneous reproduction number (R(t)) estimates obtained from our dataset. A time window of 7-days is set. Additionally, reported are the standard deviation of R(t), the 2.5th percentile and the 97.5th percentile which represent the 95% central range of R(t).

Similarly, estimates of R(t) are shown in Table 6 however, here we define a 10-day time window. The R(t) estimates range from 1.24 to 3.33. As each week progresses in time, both R(t) and the standard deviation of R(t) decreases as expected.

Table 6

Time Period	Mean R(t)	SD R(t)	2.5th Percentile	97.5th Percentile
Instantaneous Reproduction Numbers R(t) **mean SI = 3.6; SD SI = 1.6**				
4/15/09 - 4/25/09	3.33	0.16	3.04	3.65
4/16/09 - 4/26/09	3.11	0.13	2.87	3.36
4/17/09 - 4/27/09	2.59	0.10	2.41	2.78
4/18/09 - 4/28/09	2.15	0.07	2.01	2.29
4/19/09 - 4/29/09	1.80	0.05	1.69	1.91
4/20/09 – 4/30/09	1.60	0.04	1.52	1.69
4/21/09 – 5/1/09	1.48	0.04	1.40	1.55
4/22/09 – 5/2/09	1.35	0.03	1.28	1.41
4/23/09 – 5/3/09	1.24	0.03	1.18	1.29

Instantaneous reproduction number (R(t)) estimates obtained from our dataset. A time window of 10 days is set. Additionally, reported are the standard deviation of R(t), the 2.5th percentile and the 97.5th percentile which represent the 95% central range of R(t).

Overall, R(t) estimates are relatively similar irrespective of the time windows which we predefined to be either 7 or 10- day periods.

CHAPTER 4: DISCUSSION

Prior to the availability of computer packages such as R0 and EpiEstim, epidemiologists would have been required to possess both an in-depth understanding of the mathematical methods utilized in epidemiological modeling as well as a certain expertise in implementing these methods as computer programs. This aspect was well-recognized by Dr. Manoj Gambhir in his seminal paper "Infectious Disease Modeling Methods as Tools for Informing Response to Novel Influenza Viruses of Unknown Pandemic Potential". [24] The availability of ready-made computer packages obviates the need for epidemiologists to be expert programmers and bridges the gap which currently exists between mathematical modelers and public health practitioners.

In addition, as the use of these packages becomes familiar and widespread, organizations such as the CDC will potentially have access to more up to date information needed to inform decision making during public health emergencies. This is because at present, the public health scientists and practitioners at the CDC rely upon outside entities and organizations to estimate these parameters which can be time-consuming and inefficient during an epidemic/pandemic outbreak.

Below, shown in Table 7 consists of all estimates computed in this thesis and estimates reported by Biggerstaff et al. and White et al. [9, 16]

As displayed in Table 3, R values computed using the ML method in the R0 package closely agree with the results White et. al. obtained in their paper. The results between the R0 package and the original White paper are similar, supporting the conclusion that the the implementation of the ML method in the R0 package closely matches the implementation in the technical paper. This is apparent with a reproduction estimate of

2.58 for the time period starting on March 28 and ending on April 28 compared to the White et al estimate of 2.59 for the same time period. [16]

Table 7.

R package: R0 **R Estimate** ML method Time period: 3/28/09- 4/28/09	2.58
White et al. [16] **R_0 Estimate** ML method Time period: 3/28/09- 4/28/09	2.59
R package: R0 **R Estimate** ML method Time period: 3/28/09- 5/3/09	1.41
R package: R0 **R Estimate** EG method Time period: 3/28/09- 5/3/09	1.88
R package: R0 **R Estimate** TD method Time period: 3/28/09- 5/3/09	1.91
R package: R0 **R Estimate** SB method Time period: 3/28/09- 5/3/09	1.24
R package: EpiEstim **R(t) Estimate** Time window: 7 days	1.04 – 3.37
R package: EpiEstim **R(t) Estimate** Time window: 10 days	1.24 – 3.33
Biggerstaff et al. [9] **R estimate**	1.46

Comparison table of all estimates computed in this thesis and estimates found in the literature.

These estimates of 2.58 and 2.59 [16] in comparison to other estimates found, which are below 2 seen in Table 7, can be attributed to the set time period which was implemented. For instance, once we expand this time period to end on May 3, we compute a R value of 1.41 which is substantially lower and more closely related to the R value of 1.46 found in the Biggerstaff et al. meta-analysis.

Shown in Table 4 and Table 7, the estimated values of the R value by the four methods are quite apparent. The SB method which continuously updates the posterior distribution each day from the prior the day before, provides the lowest R value,1.24, This could be attributed to the fact that by doing so this method is implicitly incorporating the impact of real-time intervention. The ML method provides the next lowest estimate since it does not restrict itself to the exponential growth period of the disease as the two other methods do. The TD and the EG methods provide the highest estimates since they focus upon the period of exponential growth. Nevertheless, the median value of 1.63 for the four methods is quite comparable to the value of 1.46 obtained by Biggerstaff et. al. in their meta- analysis. During an emergency, it is recommended that the CDC estimate R values utilizing all four methods implemented in the R0 package. This is since each estimate depends on the method implemented and variation is inevitable. Furthermore, such small differences, nevertheless can lead to large variations in the assessment of required efficacy in interventions. [14] This can easily be done by using the Estimate.R function which enables the user of the R0 package to estimate the reproduction numbers for one incidence dataset using several methods. [14] This code can be seen in the appendix.

The R(t) estimates through the use of the EpiEstim package, obtained for both the 7-day and 10-day time windows somewhat parallel one another except for the fact that R(t) computed for the weekly time period decreases more rapidly therefore, reaching a lower minimum. These R(t) estimates using the Bayesian statistical method although, not exactly comparable to the results found by Sugimoto et al. (R(t): 1.4-3.3) [11] (and Lessler et al. (R(t)= 1.3-2.0) [12] since both studies take place in a camp based setting and school setting respectively and not a large scaled population setting nevertheless, are quite similar. This is especially true in comparison to the Sugimoto et al. estimates. Finally, it is important to note, that the estimation of R(t) provides important insight into the temporal changes in transmission of a disease during a pandemic. However, the interpretation of these time varying trends is not always straightforward. [13] These changes in the reproduction number, as demonstrated, can be due to decreased susceptibility in the population, changes in transmissibility, and changes in contact patterns within the affected population among other possibilities. [13]

It has been shown in the risk assessment produced by Reed et al. of the CDC, that estimates of R, which were computed through the use of the R0 package in this thesis, that are between 1.0- 1.7 are classified as having low to moderate transmissibility of influenza. [25] Further, R greater than or equal to 1.8 is considered as having moderate to high transmissibility potential. The estimates computed by the implementation of the ML (R0 = 1.35) and SB (R0 = 1.16) methods shows this pandemic to potentially have low to moderate severity. [25] In contrast, the estimates computed by the EG (R0 = 1.90) and TD (R0 = 1.91) methods shows that this pandemic could have moderate to high severity. [25] Overall, if we take the median of all four estimates we compute an R value of 1.63.

this estimate and the estimate found in the Biggerstaff et al. meta-analysis places this pandemic to potentially having low to moderate severity at such an early phase of this pandemic. [25] This is comparable to what Reed et al. report in which the 2009 pandemic was classified as moderate to low severity as well. [25] It is important to note that the various estimates presented in this thesis vary between mostly having low to moderate transmissibility and some occurring on the lower spectrum of having moderate to high transmissibility. Thus, it is important, during an epidemic, to compute these estimates using several methods in order to make comparisons to ensure proper conclusions are made.

Also, based on the risk assessment by Reed et al., a transmissibility scale from 1 to 5 was described in which a value of 1 was defined as having a reproduction number less than or equal to 1.1 and a value of 5 was defined as having a reproduction number greater than or equal to 1.8. [25]

With a median of 1.63 computed from all four estimates implemented in the R0 package, the value on the transmissibility scale is 4 which results in a symptomatic attack rate of 21-24% of the total population, 31-35% in a school setting, and 21-24% at a workplace setting. [25] In contrast, the estimate reported by Biggerstaff et al. [25] of 1.46 has a value of 3 on the transmissibility scale. This in turn results in a symptomatic attack rate of 16-20% of the total population, 21-25% in a school setting, and 16-20% at a workplace setting. As the attack rate increases, the quicker it is to reach the peak number of cases as a result of such diseases. Furthermore, these results in comparison are quite similar.

Knowledge such as this can immensely aid public health practitioners in having the ability to make decisions during early phase pandemic planning in order to quickly mitigate against such diseases. For instance, for low to moderate severity, the CDC recommends voluntary home isolation of ill persons, selective school dismissals in facilities with children at high risk for severe influenza complications, and recommendation for vaccination which is considered the main tool in reducing the risk of acquiring infection. [26]

Limitations and Observations

All methods described in this thesis, are fundamentally dependent upon the quality of the onset and reported data time series collected by organizations such as the CDC. In this study, which used the latest available data for the 2009 Influenza Pandemic in the United States, the dataset is still deficient in that not all onset dates had a corresponding report date and vice versa. Despite this however, it needs to be stressed that it does not detract from the overall value of neither this study or the ones performed by White et al.

Before performing any study and using these R packages, it is essential to scrutinize the data and undertake any modifications necessary. Since onset times are of interest, a more accurate set of incidence time series could be developed by imputing onset times by examining report delays and then based upon that adjusting incidence numbers which do not have onset dates. In addition, changes in the reporting rate of cases can significantly impact estimated if it is not constant over the entire outbreak. This can lead to underestimation of Rif the reporting rate decreases.

Another major factor which fundamentally impacts the study is the knowledge of the mean serial interval and standard deviation. These in turn determine the serial interval distribution. The assumed PDF of the serial distribution can also impact the results obtained.

For example, the serial interval mean was estimated to be 3.6 days however, if this estimate is changed, this will significantly impact the resulting reproduction number. It can be observed that a lower mean for the serial interval which in turn affects the serial interval distribution has the effect of dramatic reduction in reproduction number.

For the Bayesian statistical method another factor that impacts the results is the magnitude of time window chosen. Generally, a larger time window will smooth out the estimates however, a smaller time window can provide more frequent updates of the posterior distribution. (refer to table in appendix)

The initial prior distribution along with the likelihood function are of great importance in the Bayesian statistical method since it determines the subsequent trajectory of calculations. The more informative the prior distribution, then more likely it is to glean insight into the dynamic of the disease.

Conclusion

This study has both demonstrated the use of and further validated the two computer packages that are now available for general use by non-modelers. Although the use of these packages still requires a certain minimum knowledge of statistical methods, the availability of these packages vastly improves the tools now at the disposal of public health practitioners during an epidemic/pandemic. Further studies should be performed to both validate and demonstrate the use of these models for other diseases as well as for other datasets.

Work Cited:

1. 2009 swine flu pandemic originated in Mexico, researchers discover. *ScienceDaily*. 2016;(https://www.sciencedaily.com/releases/2016/06/160627160935.htm). (Accessed February, 2018)
2. World now at the start of 2009 influenza pandemic. *World Health Organization*. 2010;(http://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/). (Accessed February, 2018)
3. Shrestha SS, Swerdlow DL, Borse RH, et al. Estimating the burden of 2009 pandemic influenza A (H1N1) in the United States (April 2009 – January, 2010). Clin Infect Diseases
4. White LF, Wallinga J, Finelli L, et al. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza and Other Respiratory Viruses*. 2009;3(6):267–276.
5. Fraser C, Riley S, Anderson RM, et al. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences*. 2004;101(16):6146–6151.
6. Vink MA, Bootsma MCJ, Wallinga J. Serial Intervals of Respiratory Infectious Diseases: A Systematic Review and Analysis. *American Journal of Epidemiology*. 2014;180(9):865–875.
7. Cowling BJ, Fang VJ, Riley S, et al. Estimation of the Serial Interval of Influenza. *Epidemiology*. 2009;20(3):344–347.
8. Dietz K. The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research*. 1993;2(1):23–41.
9. Biggerstaff M, Cauchemez S, Reed C, et al. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC Infectious Diseases*. 2014;14(1).
10. Jones, J. H. (2007, May 1). Notes On R0. Retrieved from Notes On R0
11. Fraser C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. *PLoS ONE*. 2007;2(8).
12. Admin. Epidemic theory (effective and basic reproduction numbers, epidemic thresholds) and techniques for infectious disease data (construction and use of epidemic curves, generation numbers, exceptional reporting and identification of significant clusters). *Health Knowledge*. 2010;(https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/epidemic-theory). (Accessed February 23, 2018)
13. Cori A, Ferguson NM, Fraser C, et al. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology*. 2013;178(9):1505–1512.
14. Obadia T, Haneef R, Boëlle P-Y. The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*. 2012;12(1).
15. Spatial-R. Reproduction numbers during epidemics in R software. *Reproduction numbers during epidemics in R software | Spatial-R*. (http://spatial-r.com/en/2015/07/R0/). (Accessed February 23, 2018)

16. White LF, Wallinga J, Finelli L, et al. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza and Other Respiratory Viruses*. 2009;3(6):267–276.
17. Fraser C, Riley S, Anderson RM, et al. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences*. 2004;101(16):6146–6151.
18. Sugimoto JD, Borse NN, Ta ML, et al. The Effect of Age on Transmission of 2009 Pandemic Influenza A (H1N1) in a Camp and Associated Households. *Epidemiology*. 2011;22(2):180–187.
19. Lessler J, Reich NG, Cummings DA. Outbreak of 2009 Pandemic Influenza A (H1N1) at a New York City School. *New England Journal of Medicine*. 2009;361(27):2628–2636.
20. White LF, Pagano M. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine*. 2008;27(16):2999–3016.
21. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol. 2004;160(6):509–516.
22. Wallinga, J., and M. Lipsitch. "How Generation Intervals Shape the Relationship Between Growth Rates and Reproductive Numbers." Proceedings of the Royal Society B: Biological Sciences 274, no. 1609 (2007): 599.
23. Bettencourt, L.M.A., and R.M. Ribeiro. "Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases." PLoS One 3, no. 5 (2008): e2185.
24. Gambhir M, Bozio C, Ohagan JJ, et al. Infectious Disease Modeling Methods as Tools for Informing Response to Novel Influenza Viruses of Unknown Pandemic Potential. *Clinical Infectious Diseases*. 2015;60(suppl_1).
25. Novel Framework for Assessing Epidemiologic Effects of Influenza Epidemics and Pandemics - Volume 19, Number 1-January 2013 - Emerging Infectious Disease journal - CDC. *Centers for Disease Control and Prevention*. 2012;(https://wwwnc.cdc.gov/eid/article/19/1/12-0124_article).
26. Morbidity and Mortality Weekly Report (MMWR). *Centers for Disease Control and Prevention*. 2017;(https://www.cdc.gov/mmwr/volumes/66/rr/rr6601a1.htm).

Figures:

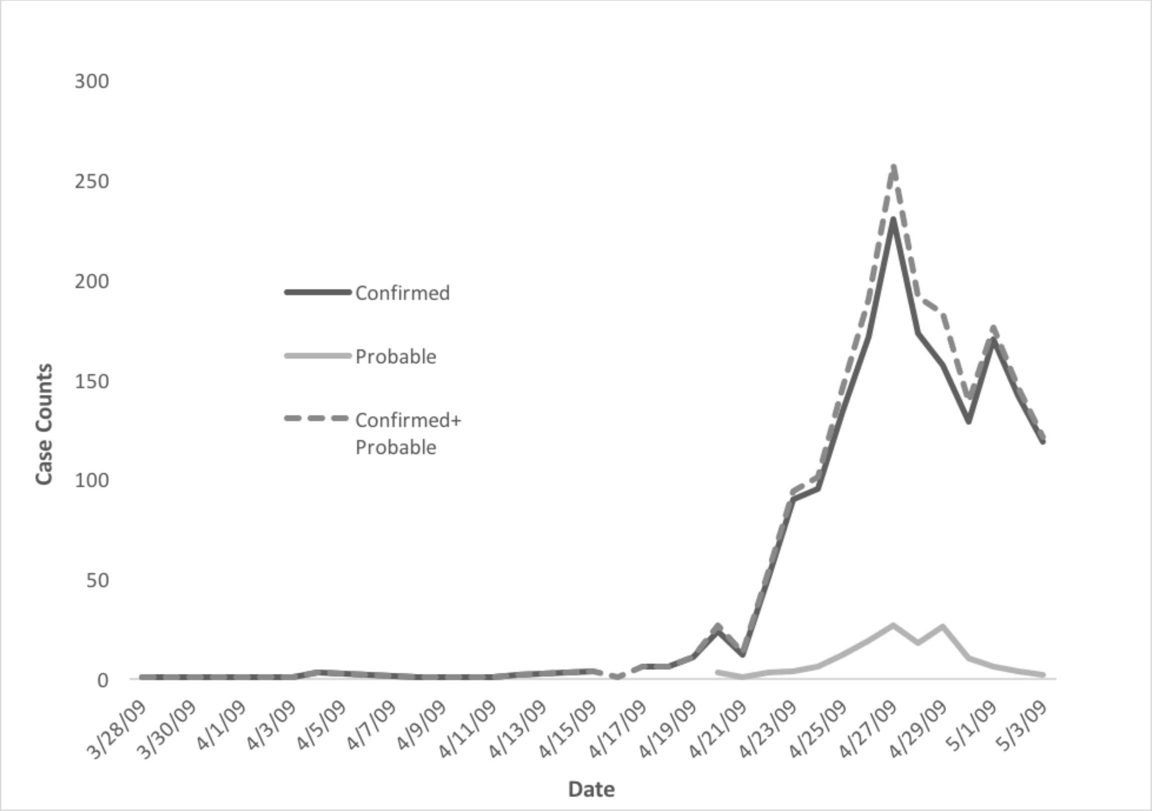

Figure 1: Confirmed and probable cases plotted by onset time occurring in the United States. First date of onset is March 28, 2009 ending on May 3, 2009. It can be observed that onset times rapidly decline as we approach the final onset date. There were 1738 confirmed cases, 142 probable cases, and a total of 1880 confirmed and probable cases.

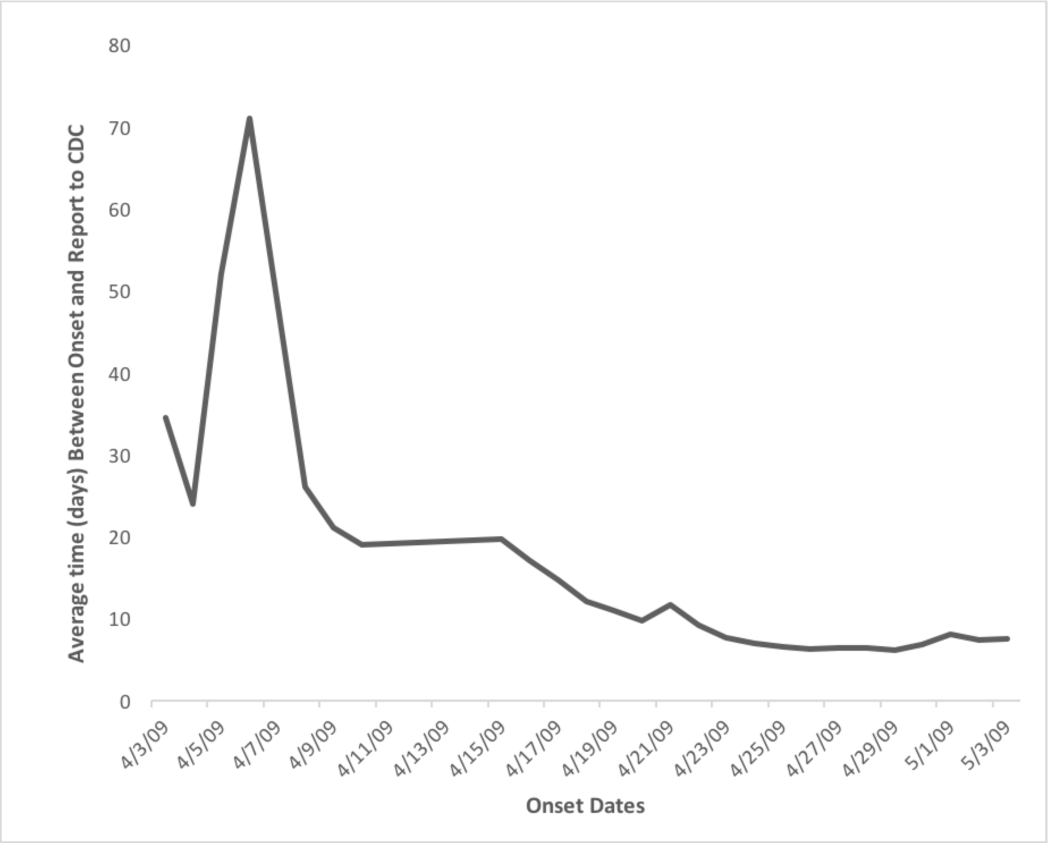

Figure 2: The average reporting delay by days to the CDC of cases with a known date of onset is shown. It can be observed that the peak reporting delay between onset and report date occurs on April 6, 2009. As time progresses, reporting delay rapidly declines.

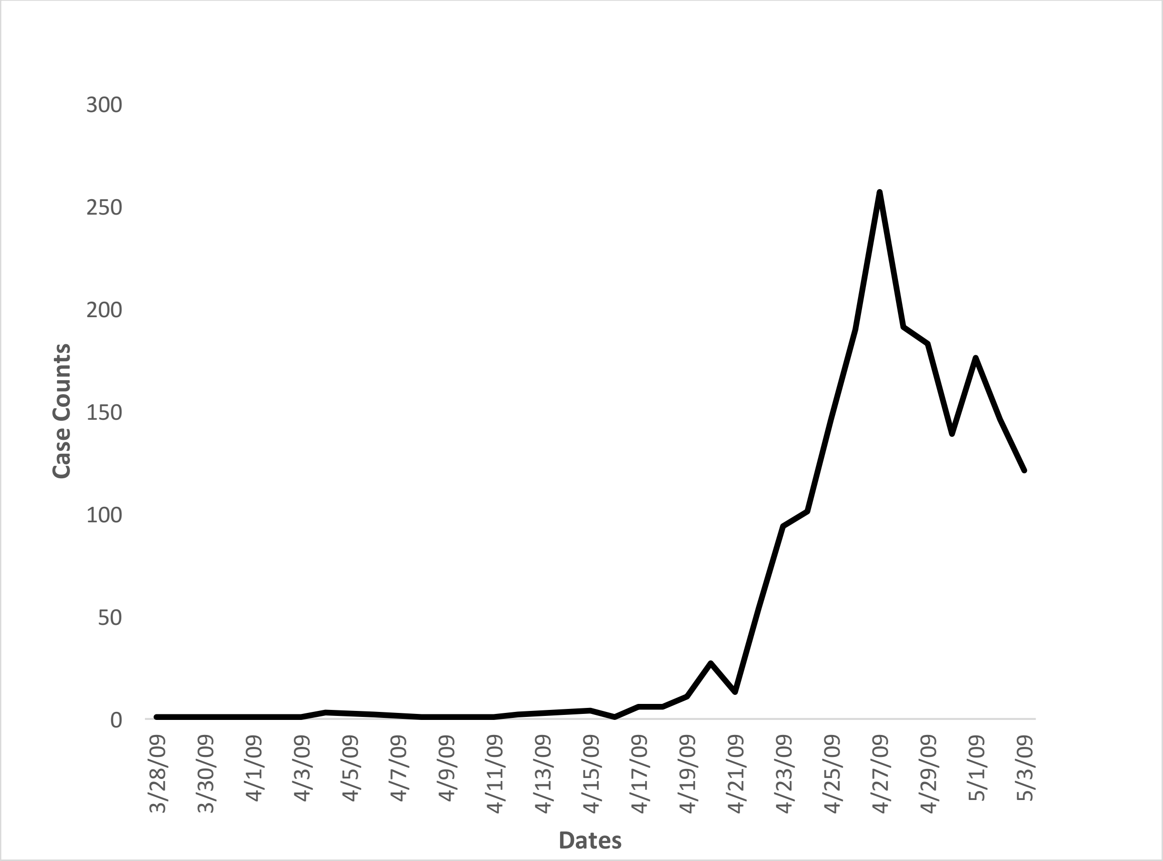

Figure 3. Epidemic curve of infection for 2009 Influenza Pandemic in the United States. The analysis begins on March 28,2009 and ends on May 3, 2009. There are a total of 1880 confirmed and probable cases with a total of 29 days of data used.

Tables:

Table 1: Included are the various parameters supplied by the R package, R0. We provide the definitions and our inputs that produce the results presented in this thesis. [14]

***R0* R package Details**		
Methods	est.R0.ML est.R0.EG est.R0.SB est.R0.TD	
Parameters	**Definitions**	**Inputs**
epid	The epidemic curve	our dataset which we call casecounts
mGT	Generation time or serial interval	Mean serial interval is set at 3.6 with standard deviation 1.6 We assume a Weibull distribution
Begin	Time estimation begins	We set this at ‘2009-3-28’
End	Time estimation ends	We set this at either: [16] '2009-4-25' '2009-4-26' '2009-4-27’ or '2009-4-28'

Table 2 Included are the various parameters supplied by the R package, EpiEstim. We provide the definitions and our inputs that produce the results presented in this thesis. [13]

EpiEstim R package Details		
Parameters	**Definitions**	**Input**
I	Vector of incidence cases	cases
T. Start, T.End	Vector of positive integers consisting of the start and end times of each time-window for with the reproduction number is estimated.	Varied (Refer to appendix for code)
Method	Method in which the serial interval distribution is specified	"UncertainSI"

n1, n2	n1- size of sample of pairs. Must be a positive integer
	n2- size of the sample drawn from each posterior distribution. Must be a positive integer.
		We set this to be:
		n1= 1000
		n2= 1000

Mean.SI	Mean serial interval	3.6
Std.SI	Standard deviation of the serial interval	1.6
Std.Mean	Standard deviation of the distribution from which the serial interval is drawn	1
Min.Mean.SI	Lower bound of serial interval distribution	3.5
Max.Mean.SI	Upper bound of serial interval distribution	3.7
Min.Std	Lower bound of standard deviation of the SI distribution	1.5
Max.Std	Upper bound of standard deviation of the SI distribution	1.7

Table 3: R estimates assume the mean of the serial interval (SI) is 3.6 days with standard deviation (σ) of 1.6 days. [7] Estimates computed by White et al. are included. Total case counts for each time period can also be seen.

Time Period	**Mean R_0**	**2.5th Percentile**	**97.5th Percentile**	**Total Cases**
Reproduction Numbers mean SI = 3.6 days; σ =1.6				
3/28/09 - 4/25/09	3.54	3.11	4.01	477
3/28/09 - 4/26/09	3.27	2.93	3.64	667
3/28/09 - 4/27/09	3.07	2.80	3.36	924
3/28/09 - 4/28/09	2.58	2.37	2.80	1115
White et al. (2009) Estimates [16]				
Original Data-Reproduction Numbers R_0 mean SI = 3.6 days; σ =1.6				
3/28/09 - 4/25/09	3.48	2.88	3.72	275
3/28/09 - 4/26/09	3.29	2.85	3.47	392
3/28/09 - 4/27/09	2.87	2.55	3.06	529
4/3/09 - 4/28/09	2.59	2.31	2.77	681

Table 4: Overall R_0 estimates assume the mean of the serial interval (SI) is 3.6 days with standard deviation (σ) of 1.6 days. [7] We implement four methods provided by the *R0* package: ML, EG, TD, and, SB. [14]

Method	**Reproduction Number** **3/28/2009 – 4/3/2009**	**95% CI**
ML	1.41	[1.32;1.50]
EG	1.88	[1.81;1.95]
TD	1.91	[1.52; 2.31]
SB	1.24	[1.08- 1.40]

Table 5: Instantaneous reproduction number (R(t)) estimates obtained from our dataset. A time window of 7-days is set. Additionally, reported are the standard deviation of R(t), the 2.5th percentile and the 97.5th percentile which represent the 95% central range of R(t**).**

Time Period	**Mean R(t)**	**SD R(t)**	**2.5th Percentile**	**97.5th Percentile**
Instantaneous Reproduction Numbers R(t) **mean SI = 3.6; SD SI = 1.6**				
4/18/09 - 4/25/09	3.37	0.16	3.07	3.69
4/19/09 - 4/26/09	3.12	0.13	2.87	3.37
4/20/09 - 4/27/09	2.55	0.09	2.37	2.74
4/21/09 - 4/28/09	2.12	0.07	1.98	2.26
4/22/09 - 4/29/09	1.73	0.05	1.63	1.84
4/23/09 – 4/30/09	1.49	0.04	1.41	1.58
4/24/09 – 5/1/09	1.32	0.04	1.25	1.39
4/25/09 – 5/2/09	1.17	0.03	1.10	1.23
4/26/09 – 5/3/09	1.04	0.03	0.98	1.09

Table 6: Instantaneous reproduction number (R(t)) estimates obtained from our dataset. A time window of 10 days is set. Additionally, reported are the standard deviation of R(t), the 2.5th percentile and the 97.5th percentile which represent the 95% central range of R(t).

Time Period	Mean R(t)	SD R(t)	2.5th Percentile	97.5th Percentile
Instantaneous Reproduction Numbers R(t) mean SI = 3.6; SD SI = 1.6				
4/15/09 - 4/25/09	3.33	0.16	3.04	3.65
4/16/09 - 4/26/09	3.11	0.13	2.87	3.36
4/17/09 - 4/27/09	2.59	0.10	2.41	2.78
4/18/09 - 4/28/09	2.15	0.07	2.01	2.29
4/19/09 - 4/29/09	1.80	0.05	1.69	1.91
4/20/09 – 4/30/09	1.60	0.04	1.52	1.69
4/21/09 – 5/1/09	1.48	0.04	1.40	1.55
4/22/09 – 5/2/09	1.35	0.03	1.28	1.41
4/23/09 – 5/3/09	1.24	0.03	1.18	1.29

Table 7: Comparison table of all estimates computed in this thesis and estimates found in the literature.

R package: R0 **R Estimate** ML method Time period: 3/28/09- 4/28/09	2.58
White et al. [16] **R_0 Estimate** ML method Time period: 3/28/09- 4/28/09	2.59
R package: R0 **R Estimate** ML method Time period: 3/28/09- 5/3/09	1.41
R package: R0 **R Estimate** EG method Time period: 3/28/09- 5/3/09	1.88
R package: R0 **R Estimate** TD method Time period: 3/28/09- 5/3/09	1.91
R package: R0 **R Estimate** SB method Time period: 3/28/09- 5/3/09	1.24
R package: EpiEstim **R(t) Estimate** Time window: 7 days	1.04 – 3.37
R package: EpiEstim **R(t) Estimate** Time window: 10 days	1.24 – 3.33
Biggerstaff et al. [9] **R estimate**	1.46

APPENDECES

Appendix A: R0 Package Code

```
library(R0)

#dates inputted
date <- as.Date(c('2009-3-28','2009-3-30','2009-4-3', '2009-4-4','2009-4-6', '2009-4-8','2009-4-9','2009-4-10','2009-4-11','2009-4-12', '2009-4-15','2009-4-16','2009-4-17','2009-4-18','2009-4-19','2009-4-20','2009-4-21','2009-4-22','2009-4-23','2009-4-24','2009-4-25','2009-4-26','2009-4-27','2009-4-28','2009-4-29','2009-4-30','2009-5-1','2009-5-2','2009-5-3'))

#case counts
casecounts<- c(1, 1, 1, 3,2,
1,1,1,1,2,4,1,6,6,11,27,13,54,94,101,146,190,257,191,183,139,176,146,121)

#concatenating vector in R
names(casecounts) <- date
#print casecounts to check
casecounts

#save
save(casecount, file="casecount.RData")
#load casecounts
load(file="casecount.RData")
```

ML Method

```
mGT<-generation.time("weibull", c(3.6, 1.6))
est.R0.ML(casecounts,mGT, begin='2009-3-28', end='2009-5-3’')
```

EG Method

```
mGT<-generation.time("weibull", c(3.6, 1.6))
est.R0.EG(casecounts, mGT, begin = '2009-3-28', end = '2009-5-3')
```

TD Method

```
mGT<-generation.time("weibull", c(3.6, 1.6))
TD <- est.R0.TD(casecounts, mGT, begin='2009-3-28', end='2009-5-3', nsim=10000)
```

SB Method

```
mGT <- generation.time("weibull", c(3.6,1.6))
SB <- est.R0.SB(casecounts, mGT)
```

<u>Estimate.R Function (Estimate R0 for one incidence dataset using several methods.)</u>

```
mGT<-generation.time("gamma", c(3.6, 1.6)) estR0<-estimate.R(casecounts, mGT,
begin='2009-3-28', end='2009-5-3', methods=c(""ML", EG",  "TD", "SB"))
```

Appendix B: EpiEstim Package Code

```
library(EpiEstim)

cases<- c(1, 1, 1, 3,2,
1,1,1,1,2,4,1,6,6,11,27,13,54,94,101,146,190,257,191,183,139,176,146,121)


#7-day Time Windows
EstimateR(cases, T.Start=14:22, T.End=21:29, method="UncertainSI",
        Mean.SI=3.6, Std.Mean.SI=1, Min.Mean.SI=3.5, Max.Mean.SI=3.7,
        Std.SI=1.6, Std.Std.SI=0.5, Min.Std.SI=1.5, Max.Std.SI=1.7,
        n1=1000, n2=1000, plot=TRUE)

#10-day Time Windows
EstimateR(cases, T.Start=11:19, T.End=21:29, method="UncertainSI",
        Mean.SI=3.6, Std.Mean.SI=1, Min.Mean.SI=3.5, Max.Mean.SI=3.7,
        Std.SI=1.6, Std.Std.SI=0.5, Min.Std.SI=1.5, Max.Std.SI=1.7,
        n1=1000, n2=1000, plot=TRUE)
```