IDENTIFIABILITY ANALYSIS OF THE HUMAN H1N1 INFLUENZA VIRUS

Vivek Sreejithkumar and Necibe Tuncer
Harriet L. Wilkes Honors College, & Charles E. Schmidt College of Science

Abstract

Influenza is a significant source of morbidity and mortality both worldwide and also in the United States. In the U.S., the Center for Disease Control (CDC) estimates over 490,000 hospitalizations and 34,000 deaths during the 2018-2019 influenza season [2]. The objective of this research is to determine the epidemiologically important parameters of the H1N1 influenza virus such as the infection and recovery rates using mathematical modeling. Publicly available influenza incidence data from the CDC webpage was used to validate the mathematical model. The spread of the H1N1 influenza virus is modeled using the Susceptible-Infected-Recovered (SIR) compartmental model. To account for vaccination and treatment of the virus, SIVR and SITR models are considered. The models were run on the computer software MATLAB to compare the predictions of the model to the CDC data. To ensure the model's precision, the parameters were manipulated so that the model predictions could mirror the data. It was found that the 2018-2019 season H1N1 influenza infection rate is 0.2567 per day and the recovery rate is 0.1774 per day. Finally, the identifiability of the models was analyzed through Monte Carlo Simulations, which were performed on MATLAB. The results show that the average relative errors of all the model parameters remained lower than the measurement errors. Thus, these results validate the identifiability of the epidemiological models considered in this study and the reliability of the parameter estimates.

Keywords: influenza virus; mathematical modeling; Kermack-McKendrick Model; parameter estimation; ordinary differential equations; epidemiology; Monte Carlo Simulation

Introduction

Background

Epidemiology is the field of science concerned with the diffusion of disease throughout human populations. Epidemiology is inherently associated to mathematics, using mathematical principles to model the manner by which infectious diseases may spread throughout a population. Infectious diseases are diseases that transmit between individuals and are caused by pathogenic microorganisms (e.g. virus or bacteria). Some examples of infectious diseases include chickenpox, HIV, West Nile virus, malaria, coronavirus, and influenza. When an infectious disease is present in a population, chance is that any contact between 2 individuals (one infected and one susceptible) could result in transmission of the disease. For example, the H1N1 influenza virus spreads through respiratory droplets that are transmitted from person to person when someone sneezes or coughs [3]. The CDC carries the epidemiological records of many infectious diseases, such as a recorded number of incidences, hospitalizations, and disease-induced deaths that resulted from the particular disease. The data collected by the CDC can be used to predict how infectious diseases will affect the population in future years, using epidemiological modeling techniques. Federal agencies and academic partners of the CDC can effectively use mathematical models, such as the Kermack-McKendrick model, to analyze the incidence data from the CDC and publish reports on the yearly activity of the influenza virus [4]. Because the CDC itself does not always conduct modeling of the data that they collect, there is a gap in knowledge and a need for predictive measures so that the public can understand how to prepare for future seasons of infectious diseases. Mathematical models, such as the Kermack-McKendrick SIR model, define parameters based on the epidemiological characteristics of infectious diseases and simulate how an infectious disease will diffuse across a group of individuals. By modifying a typical SIR epidemiological model, one can implement possible vaccination and treatment scenarios in the representation of the disease spread. In this research, various mathematical models were used to accurately represent the weekly new influenza incidences during the 2018-2019 season as reported by CDC. Furthermore, this project sought to estimate the parameter values of the systems of ordinary differential equations (Kermack-McKendrick models) that represented the spread of the H1N1 influenza virus during the 2018-2019 season and determine the identifiability of these parameter estimations.

Mathematical models have been used as an important tool in designing prevention strategies and control measures for infectious diseases. In such studies, the data reported by government health agencies is linked to mathematical models through parameter estimation. Parameters are estimated by minimizing the differences between the model predictions and the data. However, it is crucial to first analyze whether the parameter estimation problem is well posed [7]. That is, it needs to be understood whether it is possible to uniquely determine the parameters of the model from the reported data. Lack of such identifiability analysis might result in incorrect parameter values and, as a consequence, misleading strategies for prevention and control. In this research, identifiability analysis is

performed on the H1N1 influenza models (SIR, SIVR, and SITR) using Monte Carlo Simulations.

Variable	Definition
N(t)	Size of total population
S(t)	Number of susceptible individuals at time t
I(t)	Number of infected individuals at time t
R(t)	Number of recovered individuals at time t
V(t)	Number of vaccinated individuals at time t
T(t)	Number of treated individuals at time t

Table 1. *Definition of variables in the epidemiological outbreak models (1), (2), and (3)*

Parameter	Definition
β	Rate of infection
α	Rate of recovery
Ψ	Number of newly vaccinated individuals per unit of time
δ_1	Reduced rate of infection for individuals in the vaccinated class
δ_2	Reduced rate of transmission for individuals in the treated class
γ	Percent of infected individuals who undergo treatment

Table 2. *Definition of parameters in the epidemiological outbreak models (1), (2), and (3)*

Kermack-McKendrick (SIR) Model

The Kermack-McKendrick Model is one of the earliest mathematical models of infectious diseases, from 1927, representing the spread of a typical infectious disease in a constant population [5]. The typical Kermack-McKendrick Model assumes a population that consists of individuals who are susceptible (S), infected (I), and recovered (R). Let N denote the total population size, then . The epidemiological model consists of Ordinary Differential Equations (ODEs) which describes the dynamics of each class, S(t), I(t), and R(t). The derivative of the susceptible class is equal to the number of individuals who are getting infected per unit of time, multiplied by -1 since the size of the susceptible class decreases as more people become infected. Incidences are defined as the number of people who are infected per unit of time. To model incidences, first a single individual is considered. If c represents the number of contacts one infected person makes per unit of time and p represents the probability that a contact with a susceptible individual will result in transmission of the disease, then pcS/N will represent the number of new infections per unit of time by one infected individual. Using β, an epidemiological parameter representing the rate of transmission, in place of pc, the expression is rewritten as βSI/N to represent the number of new infections per unit of time. Thus, the transmission rate β is the product of the number of contacts per unit time and the probability that this contact results in transmission of the infection. Assuming that during an influenza season the total population remains constant, the model takes the following form, where β represents the transmission rate and α represents chance of recovery.

$$
\begin{aligned}
S' &= -\beta SI/N \\
I' &= \beta SI/N - \alpha I \qquad (1) \\
R' &= \alpha I
\end{aligned}
$$

From the model, it is observed that $N' = S' + I' + R' = 0$. This total population N(t) is constant, and that constant is denoted as N(t) = N. The flowchart representing the SIR model is presented in Figure 1.

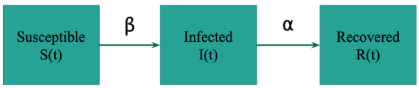


Figure 1. SIR Model – Flowchart: A methodological flowchart that demonstrates the movement of individuals between the 3 classes of the SIR model is above, where β is the transmission rate and α is the recovery rate.

SIVR Model

Many people choose to take precautions in order to avoid contracting infectious diseases such as the influenza virus. One example of such precaution is vaccination. Vaccination can be incorporated in the epidemiological model by adding a vaccinated class of individuals to the SIR Model, and thus our model becomes the Susceptible-Infected-Vaccinated-Recovered (SIVR) model [8]. Susceptible individuals get vaccinated at rate Ψ and move to the vaccinated class. Individuals in the vaccinated class become infected at a reduced infection rate of δ1 where $0 \leq \delta 1 \leq 1$. Let V(t) denote the number of vaccinated individuals, and then the final vaccination model becomes the following.

$$
\begin{aligned}
S' &= -\beta SI/N - \psi S \\
V' &= \psi S - \beta \delta_1 VI/N \\
I' &= \beta SI/N + \beta \delta_1 VI/N - \alpha I \qquad (2) \\
R' &= \alpha I
\end{aligned}
$$

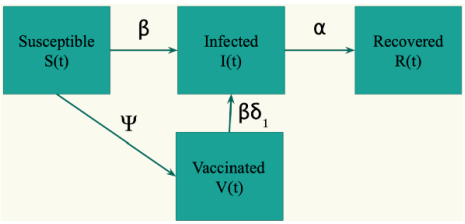


Figure 2. SIVR Model – Flowchart: A methodological flowchart showing the movement of individuals between the 4 classes of the SIVR model (Susceptible, Infected, Vaccinated, and Recovered).

Reproduction Number

A secondary epidemiological parameter that can be determined by estimating the values of the parameters in the SIR model is the reproduction number. The reproduction number of an infectious disease, also denoted as R0, is used to represent the amount of consequent infections one infected individual will cause in a fully susceptible population during his/her infectiousness period. In this type of epidemiological modeling, the reproduction number can be given by the formula R0 = β/ . The value of the reproduction number suggests certain characteristics about the disease, such as whether the disease will die out quickly or remain endemic in the population. If the reproduction number is less than one, R0 < 1, then the disease dies out in the population. If the reproduction number is greater than one, R0 > 1, then the outbreak occurs.

Methods

For this project, publicly available H1N1 incidence data is obtained from the CDC website [4]. There was no physical data collection in this project. The data used in this project consists of the number of influenza-positive tests reported to the CDC each week during the 2018-2019 season, for a whole year (52 weeks) starting from September 28th. This incidence data does not include personal information such as names. The computer software (MATLAB) is used to simulate the SIR, SIVR, and SITR models and to compare the predictions of the model to the CDC data. The incidence data was compared to the predictions of the model to match the values of the model to the data as closely as possible.

In compact form the epidemiological models (1), (2), and (3) can be rewritten as

$$\mathbf{x}'(t) = f(\mathbf{x}, \mathbf{p}) \qquad \mathbf{x}(0) = \mathbf{x}_0$$

Where x(t) is a vector of state variables, $f(\mathbf{x},\mathbf{p})$: $\mathbb{R}^3 \to \mathbb{R}^3$ is a continuous vector valued function, and $\mathbf{p} = [\beta,\alpha]$ are the model parameters. For example, for model (1),

$$x(t) = (S(t), I(t), R(t)), f(\mathbf{x}, p) = (-\beta SI/N, \beta SI/N - \alpha I, \alpha I)$$

The observations, the data collected by CDC, are functions of the state variables. Hence let y(t) denote the observations, then

$$y(t) = g(\mathbf{x}(t), \mathbf{p})$$

For model (1), the observations that is the incidences are

$$y(t) = g_1(x(t), p) = \beta SI/N$$

For model (2),

$$y(t) = g_2(x(t), p) = \beta SI/N + \beta\delta VI/N$$

For model (3),

$$y(t) = g_3(x(t), p) = \gamma I$$

Clearly, while collecting the data, the observations are contaminated with error which is called measurement errors. That means that the data do not fall on the smooth curve given by the observations y(t) above, and deviates from it. The statistical model is written as

$$y_i = g(x(t_i), p) + E_i,$$

where E_i are the measurement errors [1]. Then, to curve fit the epidemiological models (1), (2), and (3) to the CDC data, the sum of the squared differences between the measurements and the model predictions is minimized. That is, the following optimization problem is solved.

$$\hat{p} = \left(\min_{p}\right) \sum_{i=1}^{n} (y_i - g(\mathbf{x}(t_i), p))^2 \tag{4}$$

The computer software MATLAB is used to minimize (4) with the function fminsearchbnd, which is an optimization function directed to minimize this parameter estimation problem [6]. Each system of ODEs is solved using the built-in function ode15s to solve the epidemiological models (1), (2), and (3). The total population is fixed to 215,000. After fitting the mathematical model to the data, the parameter values that resulted in the best fit were recorded in results. This fitting process was repeated for each of the mathematical models (SIR, SIVR, and SITR).

The second part of this project is to analyze the identifiability of the models used to represent the spread of the H1N1 influenza virus using Monte Carlo simulations. The objective is to study whether the parameter estimation problem is well posed, in the sense that the solution is continuous. That is, if the data is varied, measuring how much the estimated parameter values will change. To achieve that, error was introduced to the model predictions at the data points gradually 1,000 times. After introducing error, the model was re-fitted 1,000 times to the data with error to observe the effect on the parameter estimates, which determines whether the initial parameter estimates are reliable. The process of Monte Carlo Simulations was executed for each of the three models, to observe the average relative errors of all the epidemiologically important parameters in all three models and measure the identifiability of the models [10]. The total average relative errors for each parameter in the respective models are also recorded as results. The Monte Carlo simulations executed in this project are itemized in detail as following.

Monte Carlo Simulation

After curve-fitting the model to the incidences data, the next step is to analyze the identifiability of the model by performing Monte Carlo Simulations. The purpose of Monte Carlo simulations is to introduce error into the incidence data to see how the parameter values of β and react [9]. To observe how the parameters change with error at each noise level, the model is re-fitted 1,000 times for each error level to conclude whether the parameter estimations are reliable. In this project, noise levels of 1%, 5%, 10%, and 20% were introduced as part of the Monte Carlo Simulations. Monte Carlo Simulations were performed in this project in the following steps.

1. Find the solutions to the epidemiological models using the true parameters $\hat{p}$ and get the output vector $g(\mathbf{x}(t), \hat{\mathbf{p}})$ at the data time points (each week).
2. Produce 1000 synthetic data sets from the statistical model that contains a set measurement error in the data to obtain the following statistical model where E_i is the average relative error introduced into the data.

$$y_i = g(\mathbf{x}(t), \hat{\mathbf{p}}) + E_i$$

3. Fit the system of ODE model according to each of the 1000 synthetic datasets created in the previous step to estimate the parameters $\mathbf{p}_j = [\beta, \alpha]$ for j = 1, 2, …1000.
4. Find the average relative errors (AREs) in parameter estimates in the set p as below

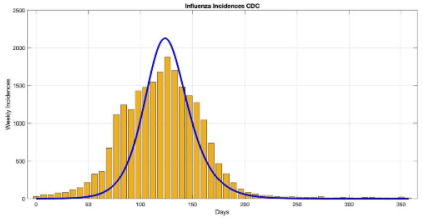
$$ARE(p^{(k)}) = 100\% \frac{1}{M} \sum_{j=0}^{M} \frac{\left| \hat{\mathbf{p}}^{(\mathbf{k})} - \mathbf{p}_j^{(\mathbf{k})} \right|}{\hat{\mathbf{p}}^{(\mathbf{k})}}$$

where $p^{(k)}$ is the kth parameter in set p, $p_j^{(k)}$ is the kth parameter in the set p_j, and $\hat{p}^{(k)}$ is the kth parameter in the set $\hat{\mathbf{p}}$ of true parameters.

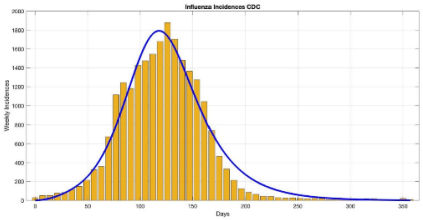
The 4 steps of the Monte Carlo simulations are repeated with each increasing level of noise and executed for each of the 3 epidemiological models (SIR, SIVR, and SITR).

Results

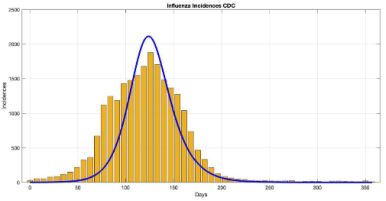
This project found success in utilizing accurate epidemiological models to represent the spread of the H1N1 influenza virus during the 2018-2019 season according to data from the CDC [4]. The value of N was fixed to 215,000 in each model to account for individuals who are infected with the disease but are not recorded by the CDC. Figures 4, 5, and 6, (below) show the curve-fitting of the mathematical model to the incidence data. The blue line represents the respective mathematical model prediction while the orange bars represent the weekly influenza incidence data. The parameters that produced the best fit to the CDC data in each model (SIR, SIVR, and SITR) are summarized in Table 3, 4, and 5 below, respectively. For example, when curve-fitting the CDC data to the standard SIR model, the estimated transmission rate (β) was 0.2567 per day and the recovery rate ( ) was 0.1774 per day. These parameters represent conditions of the population and characteristics of influenza that produce a similar pattern of incidences compared to CDC data from the 2018-2019 H1N1



(Figure 4) *Graph demonstrating curve-fitting results of the SIR Model* (1)



(Figure 5) *Graph demonstrating curve-fitting results of the SIVR Model (2)*



(Figure 6) *Graph demonstrating curve-fitting results of the SITR Model (3)*

β	0.2567 1/day
α	0.1774 1/day

Table 3. Curve-fitting results of the SIR model

β	0.4830 1/day
α	0.0548 1/day
Ψ	0.0793 1/day
δ_1	0.2138

Table 4. Curve-fitting results of the SIVR model

β	0.3141 1/day
α	0.0436 1/day
γ	1.1502 1/day
δ_2	0.6299
ν	0.1568 1/day

Table 5. Curve-fitting results of the SITR model

From the parameter values listed in the tables above, the secondary epidemiological parameter R0, or the reproduction number, can be determined. Using the values of β and estimated from the SIR Model, the value of R0 was measured to be approximately equal to 0.2567 divided by 0.1774, or about 1.447. Because the value of R0 is greater than 1 for the 2018-19 season of H1N1 influenza, this suggests that the disease will remain endemic in the population, which is consistent with the knowledge that the influenza virus seasonally returns to the population each year.

The results of the Monte Carlo Simulations are summarized in Table 6, 7, and 8 below. As higher noise levels were presented into the data, the average relative error in the parameter estimates slowly increased as well. However, there were no extreme average relative errors, suggesting that the models have strong identifiability and that the parameter estimates found in this study for all 3 models are reliable estimates of the actual parameter values in real life.

r (error in data)	β	α
0%	0	0
1%	0.26	0.40
5%	1.38	2.11
10%	2.77	4.24
20%	5.53	8.46

Table 6. *Monte Carlo simulation results for Model (1) SIR Model.*
Average relative errors for parameters β and α are listed for each level of error in data.

r (error in data)	β	α	Ψ	δ_1
0%	0	0	0	0
1%	0.99	0.49	1.49	0.89
5%	4.53	2.48	6.93	4.23
10%	7.95	5.10	12.56	7.93
20%	10.99	10.98	18.55	11.43

Table 7. *Monte Carlo simulation results for Model (2) SIVR Model.*
Average relative errors for parameters β and α are listed for each level of error in data.

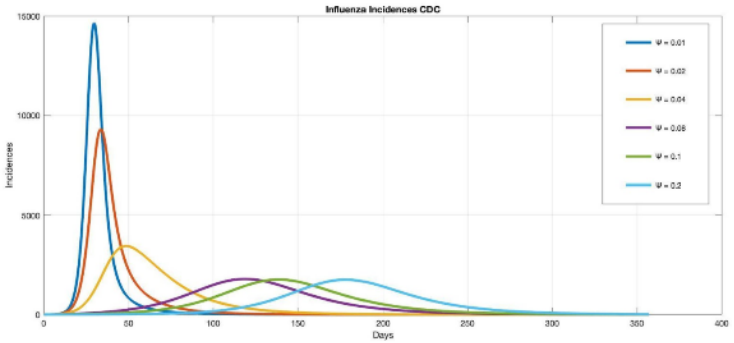
r (error in data)	β	α	γ	δ_2	ν
0%	0	0	0	0	0
1%	0.21	2.12	1.69	0.59	0.25
5%	1.09	7.26	5.67	1.77	1.63
10%	1.94	10.85	9.10	2.51	3.53
20%	3.66	17.84	17.75	4.20	7.30

Table 8. *Monte Carlo simulation results for Model (3) – SITR Model.*
Average relative errors for parameters β and α are listed for each level of error in data.

Discussion

The results show successful modeling of the reported incidences of the H1N1 influenza virus to the CDC using systems of Ordinary Differential Equation models. When running the model on the computer software, a fitted curve was achieved that corresponded to the official incidence data. Furthermore, the sum of the squared differences between the model predictions and the official data at each week was minimized with respect to the parameters of interest, meaning that the model is as close as possible to the collected data values reported by the CDC. After running Monte Carlo simulations on the models, the total average relative error found for each model was moderately low. Thus, the models have strong identifiability and the parameter estimates are reliably dependable. The epidemiological model applied in this project can be efficiently used to model the spread of various infectious diseases among a population.

This type of epidemiological modeling can be used to predict the long-term behavior of infectious diseases, such as whether a disease will die out after one season or remain in the population and become an endemic. This knowledge equips health officials to prepare for upcoming infectious disease seasons by preparing vaccines and treatments in advance, as the model can predict the number of weekly incidences of the disease. Furthermore, manipulating different parameters in the epidemiological model demonstrates how different factors influence the dynamics of the spread of the influenza virus. For example, Figure 7 (shown below) demonstrates how manipulating the vaccination rate affects the epidemiological model after the other parameters is fixed to the value obtained in this project.



(Figure 7) *Graph of showing manipulation of the vaccination rate (Ψ) in the SIVR model*

In this project, the vaccination rate was estimated to be 0.0793 or approximately 0.08 per day. Multiplying 0.08 times S (number of susceptible individuals in the population) will provide the number of people vaccinated per day. In Figure 4, the purple curve matches the vaccination rate that was estimated in this project and exhibits a number of incidences similar to the data reported by the CDC. The different curves in Figure 4 demonstrate how different vaccination rates could affect the number of influenza incidences over time. As seen in Figure 4, when the vaccination rate in the model is lowered to 0.01, the number of influenza incidences rapidly increases and reaches its peak at almost 15,000 incidences per day at about 25th day after the start of the influenza season. When the vaccination rate is increased to 0.2, it is observed that the curve has been flattened (see light blue curve). It reaches its peak at a later day, approximately 175th day after the start of the influenza season and the peak is at a lower value, 1000 incidences per day. This demonstrates how the dynamics of the epidemiological model greatly depends on the values of the various parameters. If the vaccination rate is low, then there will be a lot of new infections per day. The peak of incidences occurs later for higher vaccination rates. By showing how if less people get vaccinated, the number of incidences rapidly increases demonstrates the importance of vaccination in controlling the spread of an infectious disease.

REFERENCES

Banks H.T., Hu S., Thompson W.C. (2014). Modeling and inverse problems in the presence of uncertainty. CRC Press, Boca Raton

Center for Disease Control. (2020, January 8). Estimated Influenza Illnesses, Medical visits, Hospitalizations, and Deaths in the United States — 2018–2019 influenza season. Retrieved May 19, 2020, from https://www.cdc.gov/flu/about/burden/2018-2019.html

Centers for Disease Control. (2018, August 27). How Flu Spreads. Retrieved May 19, 2020, from https://www.cdc.gov/flu/about/disease/spread.htm

Center for Disease Control. (2019, October 15). National, Regional, and State Level Outpatient Illness and Viral Surveillance. Retrieved May 19, 2020, from https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

Martcheva, M. (2015). An introduction to mathematical epidemiology. Springer, New York.

MATLAB. (2020). Fminsearch. Retrieved May 19, 2020, from https://www.mathworks.com/help/matlab/ref/fminsearch.html

Miao H, Xia X, Perelson AS,Wu H (2011) On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM Rev 53(1):3–39

Tornatore, E., Vetro, P., and Buccellato, S. (2014), SIVR epidemic model with stochastic perturbation, Neural Computing & Applications, 24(2), 309–315. https://doi-org.ezproxy.fau.edu/10.1007/s00521-012-1225-6

Tuncer, N., and Le, T. T. (2018), Structural and practical identifiability analysis of outbreak models, Mathematical Biosciences, 299, 1–18. https://doi-org.ezproxy.fau.edu/10.1016/j.mbs.2018.02.004

Tuncer, N., Marctheva, M., LaBarre, B., & Payoute, S. (2018). Structural and Practical Identifiability Analysis of Zika Epidemiological Models. Bulletin of Mathematical Biology, 8, 2209.https://doi-org.ezproxy.fau.edu/10.1007/s11538-018-0453-z