Vaccine-Preventable Diseases: Disease Modelling

Submitted in partial fulfillment of the requirements of the degree of

Interdisciplinary Dual Degree in KCDH

By

Anuj Agrawal
180110012

Under the guidance of

Prof. Ganesh Ramakrishnan

Koita Center for Digital Health (KCDH)
Indian Institute of Technology Bombay
Mumbai 400 076

Year - 2023

Specimen `B': Approval Sheet

180110012 09-Aug-23 05:45:18 PM
Vaccine-Preventable Diseases:
Disease Modeling

This thesis/dissertation/report entitled (Title) by (Author Name) is approved for the degree of M.Tech in health Informatics (Degree details).

180110012 09-Aug-23 05:46:21 PM
Anuj Agrawal

Examiners

Digital Signature
Kshitij Jadhav (10002036)
07-Aug-23 03:23:07 PM

Prof. Kshitij Jadhav

Prof. Ashish Jha

Supervisor (s)

Digital Signature
Ganesh Ramakrishnan (i09037)
07-Aug-23 05:29:12 PM

Prof. Ganesh Ramakrishnan

Chairman

Date : 7 August 2023

Place : Mumbai

https://ams.iitb.ac.in/d/117102-M7M11DPPOGCQ1RLE

Specimen `C' – Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Digital Signature
Anuj Agrawal (180110012)
07-Aug-23 03:19:16 PM

(Signature)

180110012 07-Aug-23 03:19:26 PM
Anuj Agrawal

(Name of the student)

180110012 07-Aug-23 03:19:33 PM
180110012

(Roll No.)

Date: 180110012 07-Aug-23 03:19:47 PM
07/08/2023

https://ams.iitb.ac.in/d/117102-M7M11DPPOGCQ1RLE

Abstract

Epidemiological modeling plays a pivotal role in understanding the transmission dynamics of infectious diseases and guiding evidence-based public health interventions. This report aims to provide an overview of epidemiological modeling, its methodologies, and dive deeper into two experiments/projects in that field.

The first chapter dives deep into the field of epidemiological modeling. Starting with a simple SIR model, the chapter also touches on various other models in the SIR-model family that can be used for various modeling experiments.

The second chapter describes the Measles experiment in detail, where we apply the concepts of SIR modeling, utilizing real world data and modeling the spread of the disease in Mumbai. The experiment was especially insightful due to the recent outbreak of the Measles disease in the Mumbai sub-urban areas, affecting hundreds of children. We conclude the chapter by suggesting alternate techniques, and challenges and opportunities for better modeling.

The third chapter lists out the details of the Rubella vaccination experiment. Though largely literary in nature, This report lists out two main research papers that reflect the applications of epidemiological simulations in driving public policy, like vaccination strategies that are of immense importance for a country like India.
The last chapter concludes the report and leaves suggestions for future research.

Table of contents

Chapter 1

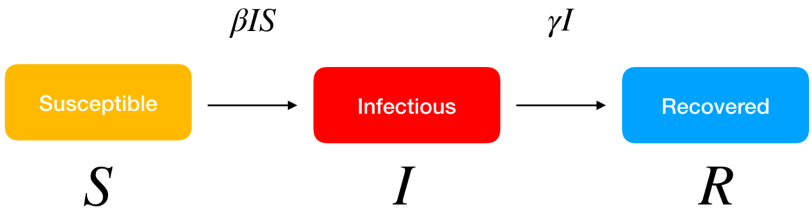
Disease Modeling: Vaccine-preventable diseases

Introduction

Vaccine-preventable diseases, like COVID-19, are of much importance to monitor and control since they have an established path of treatment. Vaccines are widely used to limit the spread of disease, and establishing a robust disease modeling and epidemiological model can help optimize vaccine delivery resources and minimize the loss of health. In this project, we undertook exploratory research to understand the various types of vaccine-preventable diseases, with a major focus on COVID-19. We explored various challenges (C.J.E.Metcalf et al.) prevalent in disease modeling research and some examples of disease modeling projects (Wadhwani AI) undertaken during the pandemic.

The SIR model

The SIR family of models is a subset of compartment models, where the population is divided into sub-groups based on their status of interaction with the pathogen. The simplest model, the SIR model refers to 3 compartments, namely - Susceptible (S), Infected (I), and Recovered/Removed (R).



This simple model can be represented by differential equations as follows:

$$\frac{dS}{dt} = -\beta IS \qquad \frac{dR}{dt} = \gamma I \qquad \frac{dI}{dt} = \beta IS - \gamma I$$

- Change of Susceptible population is negatively dependent on the existing population and infected population, with a parameter beta, the infection rate (eq - 1)
- Change in Recovered or Removed (Dead) population is only dependent on infected population, scaled via gamma, the recovery/removal rate (eq - 2)
- Change in the infected population is the summation of other two compartments, since the total population of an area is assumed to be constant (eq - 3)

System and Process Design

Models developed on data for one region often do not work on other regions. This demands that models should be trainable across multiple data sources and offer modularity and extensibility to adapt to various locales and application scenarios. These modular components are developed for independent tasks, like data ingestion, preprocessing and exploratory analysis, model fitting, scenario-conditioned forecasting, and application-specific report generation.

Modeling Methodology

While developing a model, there are various decision points that need to be considered to create a robust and scenario-based model. The loss function, optimization strategy as well as data smoothening strategies can play a huge role in the performance of the model. Data availability plays an important role in these decisions. The type and granularity of data availability, consistency, and decision-making frequency often make some model choices ideal for making public policy interventions more effective than others.

Epidemiological Model Choice

Similar to the modeling methodology, the type of epidemiological model used for modeling purposes helps in the interpretability of the model. In this research, we have majorly used the SIR family of models since they are an optimal choice with the type of data available for the COVID pandemic. It also gives out results and parameters that are easily interpretable and controlled using public interventions, giving us more control over the desired outcomes.

Interpretability and Identifiability

Certain models contain parameter choices that are not directly identifiable via the available datasets. This makes the implementation and interpretation of such models difficult and inaccurate over long periods of time. This affects the confidence of both researchers as well as policymakers in the model and hampers the effectiveness of utilizing the full potential of the model. Certain techniques, like secondary and tertiary research, data collection, and approximations are implemented to minimize this effect to improve the efficiency of the model.

Deployment and Reporting

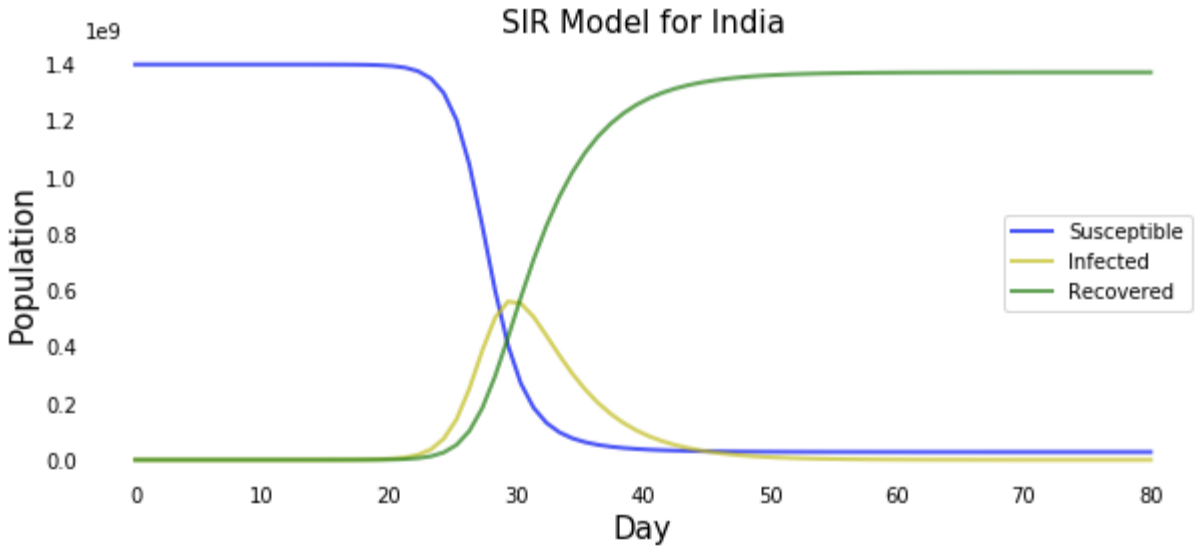
The success of any model is dependent on the reporting and generation of actionable insights to fulfill the objectives of any disease modeling exercise - reducing caseload and deaths. Therefore, it is of utmost importance that the deployment and reporting of the model are implemented according to the assumptions of the model, and the dynamic data keeps improving the performance as well as the effectiveness of the model.

SIR family models

The SIR family models are characterized by various compartments in which the population can be divided based on the interaction of an individual with the pathogen. The most basic model of this family is the SIR model, which stands for Susceptible-Infected-Recovered. Other models add a few compartments to this class, for example:

SIR model

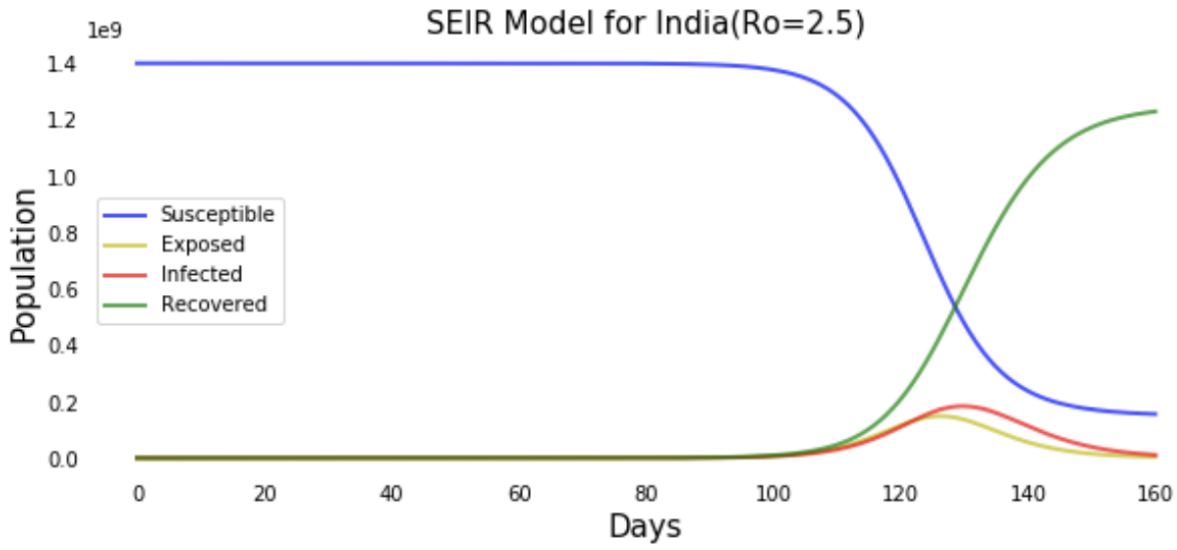
This is the most basic class of model that only includes three compartments - Susceptible (S), Infected (I), and Recovered (R). This is the simplest and an easy model to implement directly from the available dataset. However, due to the simplicity of the model, it is often not able to predict the infection rates as accurately as other advanced models. An example of an SIR model simulation on the India dataset is shown below:

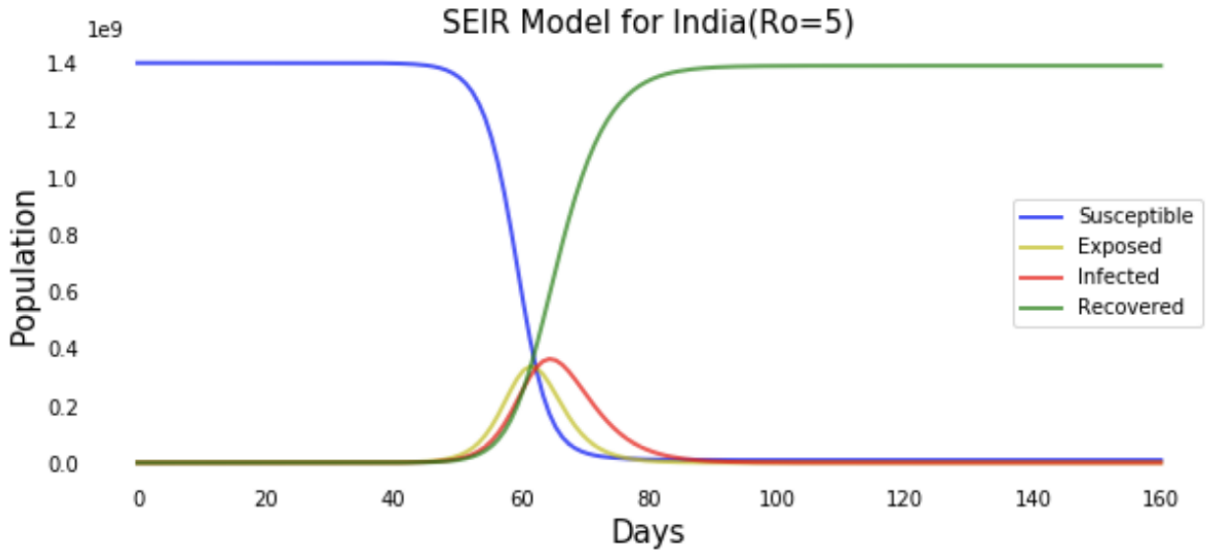


SEIR model

This adds a new compartment to the SIR model namely Exposed (E). Exposed here means that part of the population that has been in close contact with an infected individual. Adding this compartment helps the model figure out how to contain and monitor asymptomatic cases, or people who are in the hibernation mode and will show symptoms after a critical virus density is reached.

However, since the data about the number of exposed individuals is not available directly, it has to be approximated via other sources of information, like literature sources on the probability of infection once a person comes in contact with an Infected person. Some examples of the SEIR model simulation on the India dataset is shown below:

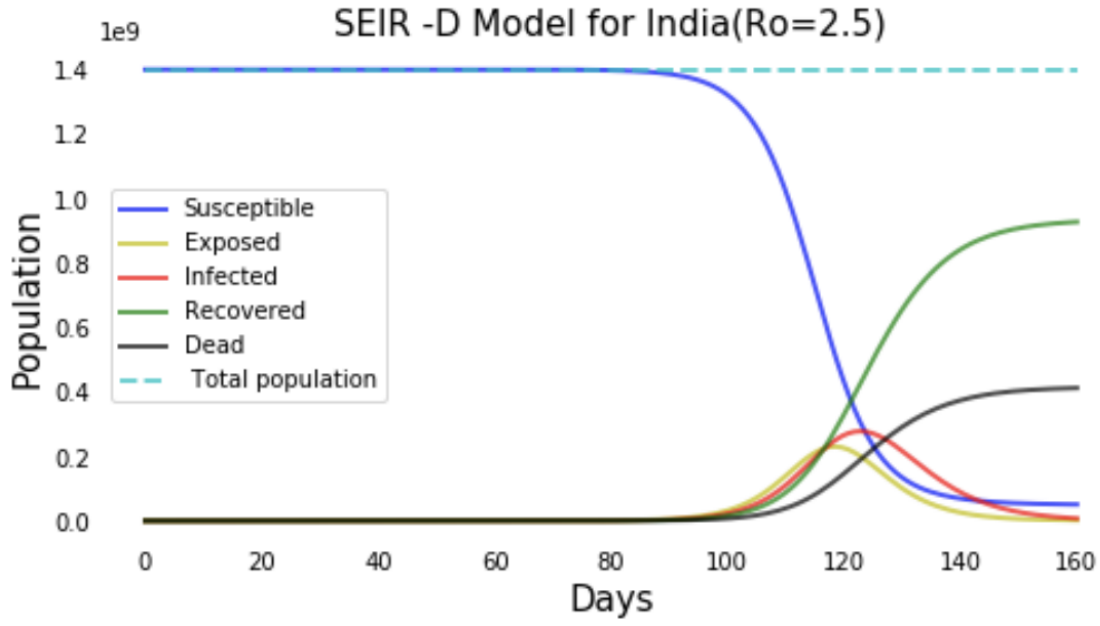


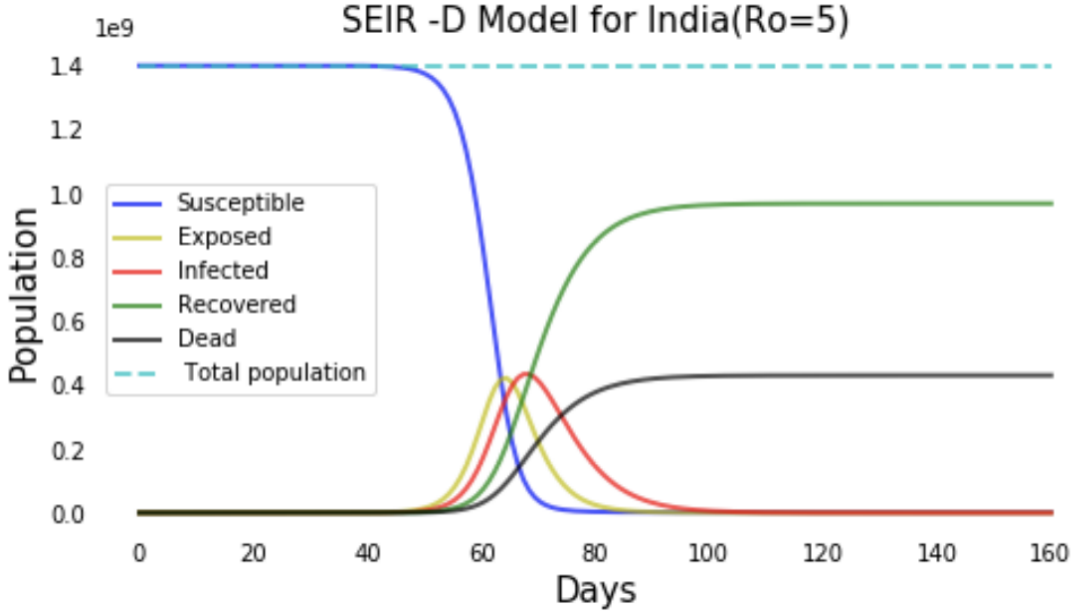


These models also help us understand the impact of R_0, the basic reproduction number, or the number of individuals infected by every infected person before they recover. It affects both the speed and intensity of a wave, and can lead to drastic outcomes based on its value.

SEIRD model

In an SEIRD model, we make a difference to separate the distinction between Recovered (R) and Dead (D) individuals. This is mainly done to measure the impact of any policy decision on the death ratio, without major impact in the infection rates. Interventions like availability of ICU beds are a good example for these models. Some examples of SEIRD model simulations on the India dataset for multiple values of R_0 are shown below:





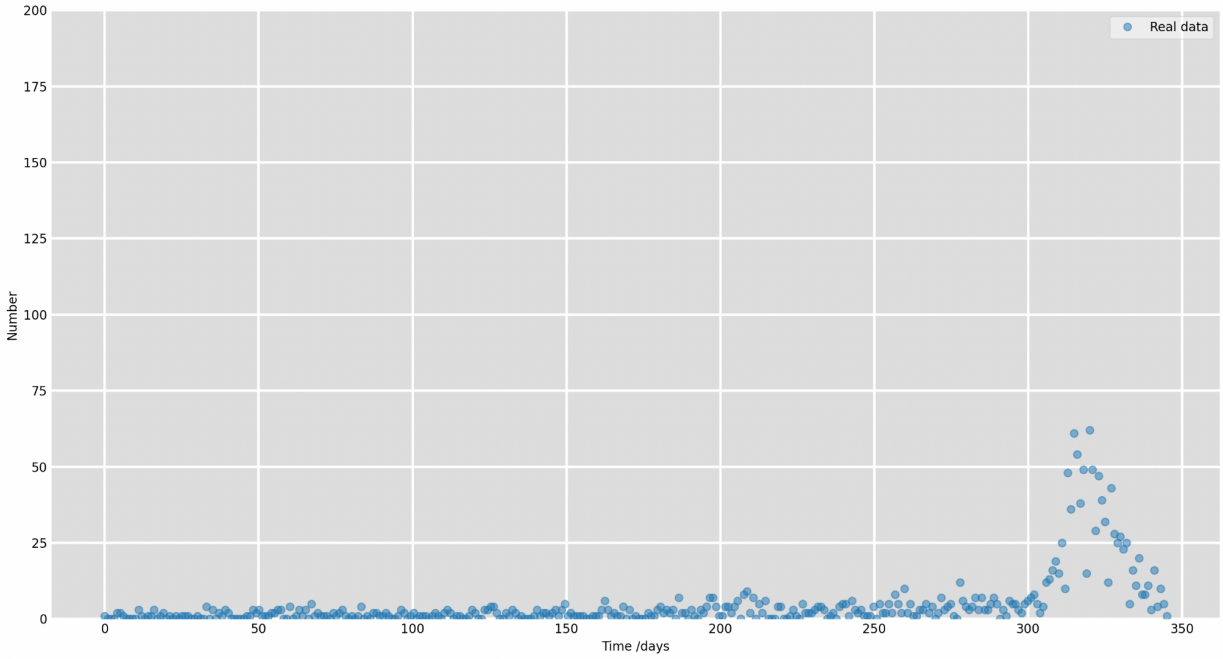
Please note that these simulations are not directly comparable to actual data, since a lot of fine-tuning and computational power was required to fit the parameters used in these models.

Chapter 2

Measles data simulation

Introduction

After much effort, we were able to get the measles data from MCGM relating to the measles outbreak in Mumbai. The data consisted of time-stamped case reports of measles cases for all wards for the entire year (Nov 21' - Nov 22'). This data helped us model the transmission of the measles disease across various wards across Mumbai and try to model the disease spread. Below is a snapshot of the data that has been processed to indicate the number of cases on a particular day. In the below graph, we clearly see a spike in cases towards Nov-2022.



Data Preparation

To further clean the data to be used for experimentation, we decided to just use 70 days worth of data near the end of the graph, so that we can capture the increase in cases and model that duration of high growth accurately. This decision was also supported by the inability of our model to accurately model the entire 350 days worth of data, due to natural asymmetries evident in the graph above. Hence, all analysis and modeling presented below are based on the last 70 days worth of data that we got from the MCGM.

SIR Modelling

$$\frac{dS}{dt} = -\beta IS$$

$$\frac{dI}{dt} = \beta IS - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

The above 3 equations depict the simplest version of an SIR model, that we used for the initial analysis of the data. Code files, along with data and all other utility functions are available at this link.

S - The number of susceptible population in the control area
I - The number of infected individuals in the population
R - The number of recovered (+ deceased) individuals in the population

The modeling consisted of two processes:

Parameter estimation

As seen from the above equations, there are mainly two parameters in an SIR model that we need to estimate to accurately predict future cases of a particular disease:

Beta (β) - This parameter tells us about the rate of infection that happens among the susceptible population.

Gamma (γ) - The gamma value tells us about the recovery rate of infected individuals.

To estimate these parameters, there are a variety of ways available in the scientific community based on the circumstances of the disease and locality:

Parameter estimation through biological evaluation

This type of parameter estimation looks at the biological properties of the pathogen, and includes analyzing the pathways through which the pathogen attacks the host. This is a more accurate measure for parameter estimation, however, it requires extensive research to accurately estimate the parameters for each pathogen, along with their variants and furthermore requires the knowledge of variant proportions in the infected population for accurate simulation.

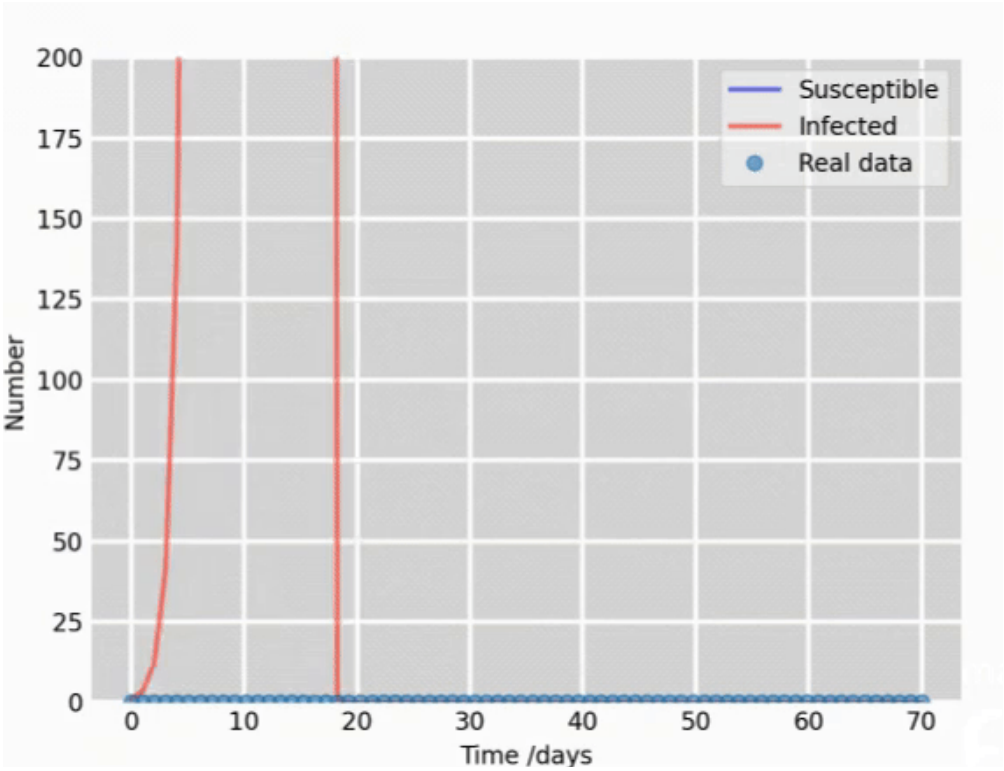
Parameter estimation through curve-fitting

This method leverages the availability of historical data to estimate parameters based on mathematical operations. There are various mathematical theorems that allow us to estimate parameters in differential equations. In our research, we used this method as it allowed us to rapidly experiment with the available data and estimate the parameters, which were then later utilized to predict future cases in the locality.

Caseload prediction

Once we have an accurate estimate of the SIR model parameters, we can just plot the transmission of active cases, recovered and susceptible population graphs and compare it with actual data to analyze the results.

The Experiment



The above animation gives us a glimpse of the final results of the experiment. Details are described below:

Process:

- For every point in the time-series data (x), we used all data points before x to estimate the parameter values
- Using these estimated parameter values, we plot the red curve to predict the future cases in the population

Parameter estimation:

- For parameter estimation, we used the least-squared-error method to minimize the squared-error of the red curve with the data points
- To perform the least-square-error method, we used the `curve_fit` method available in the scipy library in python

All code files and data are available at this link.

Optimisation

To optimize the process further, we used advanced model types like the SEIR model, that incorporates the exposed population into the analysis. For this, we assumed the ward population as the exposed population for the analysis and ran the experiment. However, the new process was not able to improve the accuracy of our simple SIR model, and hence was discarded.

Chapter 3

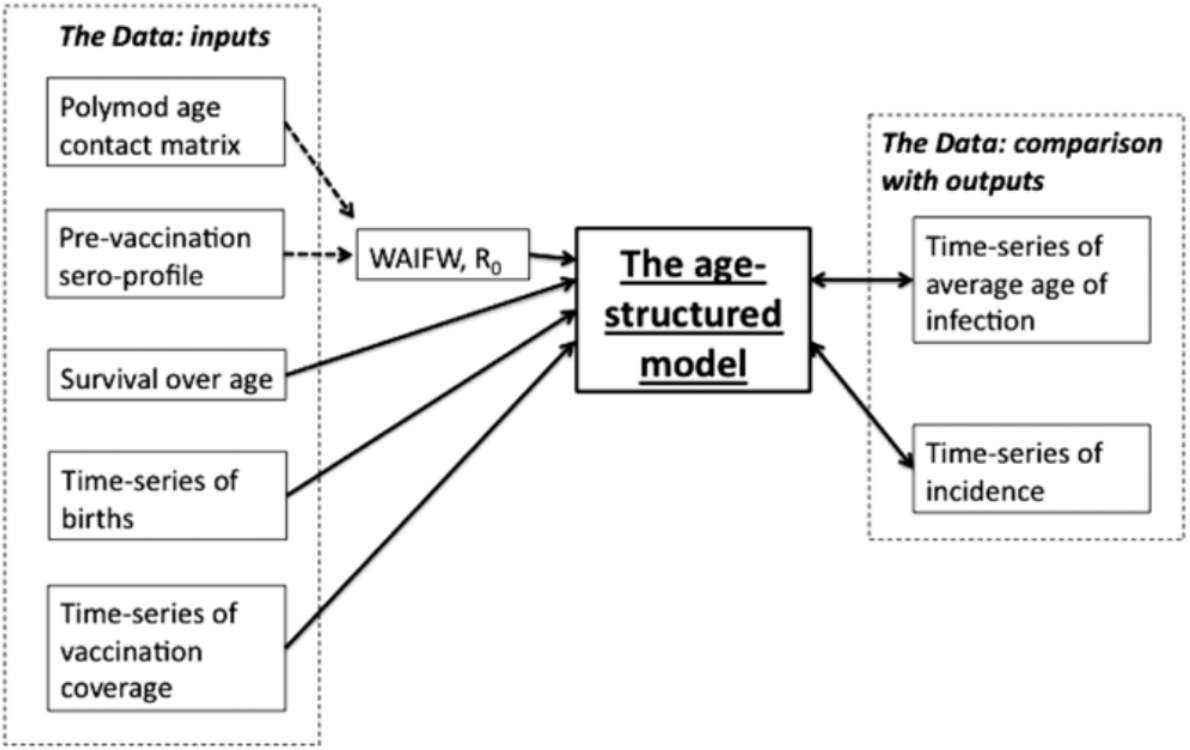
Rubella disease modeling

Introduction

Measles and Rubella have a single vaccine known as the MMR vaccine. Almost all research and epidemiological study takes the two diseases to analyze the effects of vaccination and to model disease spread. Therefore, we undertook extensive literature study and analysis about the best ways to model the diseases epidemiologically, and will be presenting some research in that field.

Structured models of infectious disease: Inference with discrete data

This study presented some of the pathbreaking research into a comprehensive study of disease epidemiology, by integrating various data sources, including birth/death data, vaccination status, contact matrices and other epidemiological parameters to create an overarching model that can be used for multiple diseases.



Details of all parameters are described below:

Polymod age contact matrix - This matrix describes the contact probabilities between various age groups in a location. This matrix was based on a study in Europe, the details of which can be found in this paper.

Pre-vaccination sero-profile - Calculates the seropositivity of a population prior to vaccination campaigns

Survival over age - The survival (or death) patterns in a population

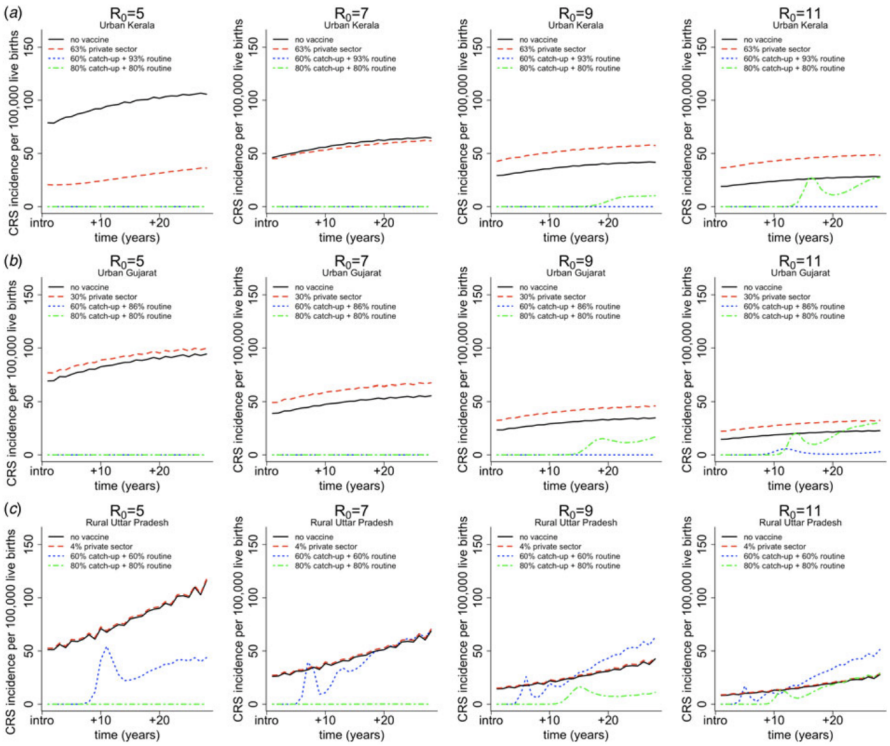
Time-series of births - Birth data in a population

Time-series of vaccination coverage - Vaccination data of the MMR vaccine to change the susceptible population.

WAIFW - Who-acquires-infection-from-whom matrix, determining the transmission of disease spread in the population.

Rubella vaccination in India: identifying broad consequences of vaccine introduction and key knowledge gaps

This paper applies the above process in the context of Rubella vaccination in India, simulating various vaccination coverage scenarios with different disease parameters (like the R0 value) to determine the optimal vaccination schedule in the country. The below figure lists out some of these scenarios, with changing R0 values and vaccination strategies.



Chapter 4

Conclusions and Future Work

With the Disease Modeling: VDP project as the DDP project, we have shown the impact and potential of the SIR family of models for simple disease modeling simulations. With various challenges and opportunities of research in the disease modeling topic, our next steps will be to identify and start working on a specific topic and research about innovative models to solve the upcoming challenges and expand the horizons of disease modeling.

Specifically, we intend to engage in more literature survey, and interact with researchers in this field to get a deeper understanding and pick up a topic for research. Given the abundance of data for COVID, we intend to keep focusing on datasets available for the disease, and explore other opportunities of work to model disease spread effectively

Finally, we can also try different model types, enriching it with more data points to create a better epidemiological approximation. Creating contact matrices in the local context can also be explored, since the current POLYMOD study reflects the societal patterns of European countries, they often fail to simulate an Indian society, leading to errors. Also by combining existing data types, we can create a better understanding of the contact and disease spread patterns and craft novel methodologies to model the population. All these areas can be explored as a continuation of this research project.

Thanks

References

- https://www.ibm.com/cloud/learn/confidential-computing
- https://sci.ldubgd.edu.ua/jspui/bitstream/123456789/9776/1/%D0%93%D0%A0%D0%90%D0%90%D0%9B%D0%AC%20%D0%9D%D0%90%D0%A3%D0%9A%D0%98.pdf#page=265
- https://www.beekeeperai.com/blog/58111-securing-healthcare-ai-with-confidential
- https://ieeexplore.ieee.org/abstract/document/9152787
- https://www.researchgate.net/profile/Dalton-Valadares/publication/358647013_Confidential_Computing_in_CloudFog-based_Internet_of_Things_Scenarios/links/620d634308bee946f3867858/Confidential-Computing-in-Cloud-Fog-based-Internet-of-Things-Scenarios.pdf
- https://link.springer.com/chapter/10.1007/978-3-030-81242-3_11
- https://forms.office.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR37R7JFLKbBAml_g6YTMEqtUN0lDSFVRTVNJUFNRNlBaNk9RMVBZMjhOSS4u
- https://confidentialcomputing.io/wp-content/uploads/sites/85/2021/02/Confidential-Computing-Consortium-Webinar-11-2020.pdf
- https://confidentialcomputing.io/projects/
- https://confidentialcomputing.io/white-papers-reports/
- https://openenclave.io/sdk/
- https://enarx.dev/
- https://research.ibm.com/blog/what-is-confidential-computing
- https://www.ijstr.org/final-print/nov2020/Mathematical-Modeling-And-Forecasting-The-Spread-Of-Covid-19-Using-Python.pdf
- https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html
- http://people.wku.edu/lily.popova.zhuhadar/
- https://github.com/SwissTPH/openmalaria/
- https://reader.elsevier.com/reader/sd/pii/S1755436514000395?token=4EB11A995983D6DE6419587BAD90D66ADF840BC0982FBF0BF45B8EEE5CFAE7F3D979F3C60B8F60253B2A98DE61419963&originRegion=eu-west-1&originCreation=20221017162732
- https://drive.google.com/file/d/17vlEbv1Ta6L6QbgFuGbBEjOhP6bvOD1f/view?usp=sharing
- https://drive.google.com/file/d/1fWN6n96wSW6WVQfmpZq9byM_5EuvViyG/view?usp=sharing
- https://drive.google.com/file/d/1nulgVUBZEgGDZpqzxRlu9epduIoD-VNE/view?usp=sharing
- https://drive.google.com/file/d/1u0DRTJOgsZrg4ZfiZ-UHjR5zYf1GfVSj/view?usp=sharing

- https://drive.google.com/file/d/19EDqVAm74T8ZOlEcSAn3D61oIRZWaMYV/view?usp=sharing
- https://github.com/lisphilar/covid19-sir
- https://covid19datahub.io/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7376536/
- https://docs.idmod.org/projects/emod-hiv/en/2.20_a/model-seir.html
- https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases
- https://github.com/Lewuathe/COVID19-SIR/
- https://github.com/PingEnLu/Time-dependent_SIR_COVID-19/
- https://www.coursera.org/specializations/infectious-disease-modelling
- https://www.coursera.org/learn/developing-the-sir-model?specialization=infectious-disease-modelling
- https://www.coursera.org/learn/interventions-and-calibration?specialization=infectious-disease-modelling
- https://www.coursera.org/learn/building-on-the-sir-model?specialization=infectious-disease-modelling
- https://github.com/AnujAgrawal30/measles_SIR_model
- https://www.cdc.gov/vaccines/vpd/mmr/public/index.html
- https://www.sciencedirect.com/science/article/abs/pii/S0040580911001031
- https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050074
- https://pubmed.ncbi.nlm.nih.gov/29198212/