

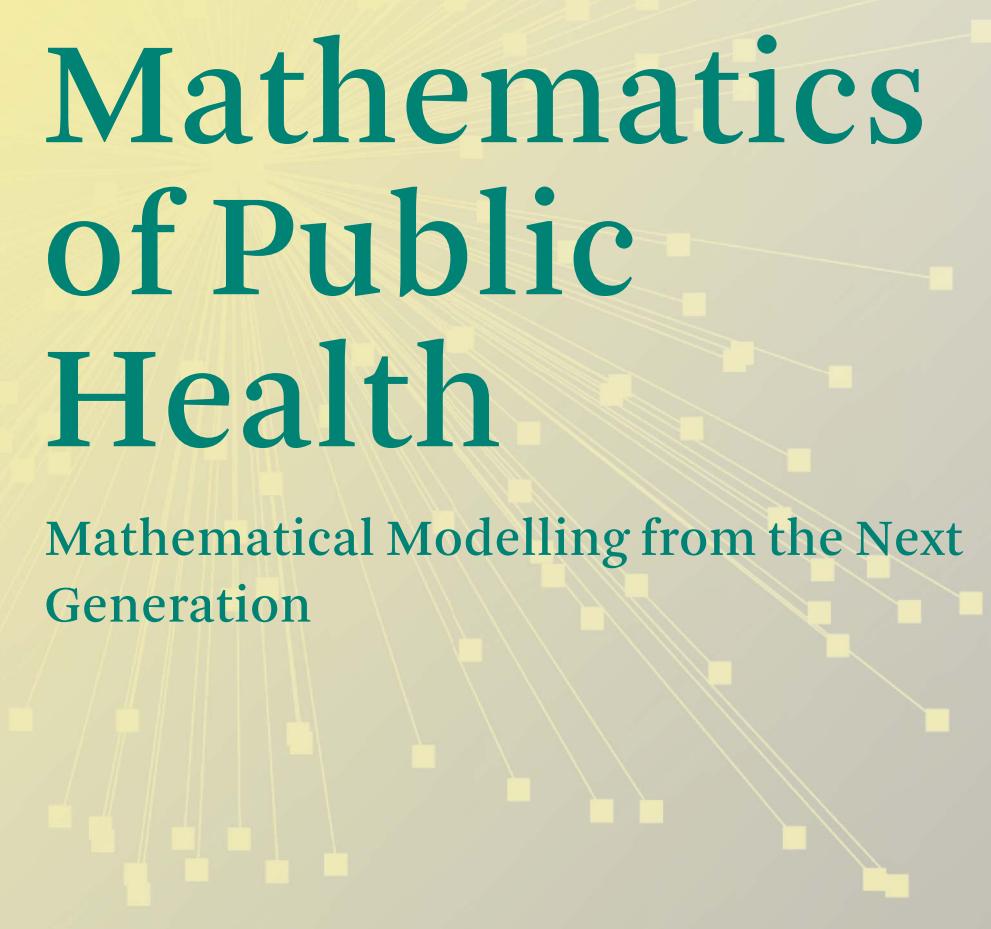
The Fields Institute for Research in Mathematical Sciences

Jummy David
Jianhong Wu
Editors



Mathematics of Public Health

Mathematical Modelling from the Next
Generation



Fields Institute Communications

Volume 88

Editorial Board Members

Deirdre Haskell, Fields Institute for Research in Mathematical Sciences, Toronto, ON, Canada

Lisa C. Jeffrey, Mathematics Department, University of Toronto, Toronto, ON, Canada

Winnie Li, Department of Mathematics, Pennsylvania State University, University Park, PA, USA

V. Kumar Murty, Fields Institute for Research in Mathematical Sciences, Toronto, ON, Canada

Ravi Vakil, Department of Mathematics, Stanford University, Stanford, CA, USA

The Communications series features conference proceedings, surveys, and lecture notes generated from the activities at the Fields Institute for Research in the Mathematical Sciences. The publications evolve from each year's main program and conferences. Many volumes are interdisciplinary in nature, covering applications of mathematics in science, engineering, medicine, industry, and finance.

Jummy David • Jianhong Wu
Editors

Mathematics of Public Health

Mathematical Modelling from the Next Generation



Editors

Jummy David
Laboratory for Industrial and Applied
Mathematics, Department of Mathematics
and Statistics
York University
Toronto, ON, Canada

Jianhong Wu
Laboratory for Industrial and Applied
Mathematics, Department of Mathematics
and Statistics
York University
Toronto, ON, Canada

ISSN 1069-5265

Fields Institute Communications

ISBN 978-3-031-40804-5

<https://doi.org/10.1007/978-3-031-40805-2>

ISSN 2194-1564 (electronic)

ISBN 978-3-031-40805-2 (eBook)

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

In early 2020, Canada was hit with one of its most significant public health crises in recent memory. The COVID-19 pandemic demanded a collaborative response from key stakeholders in health care, government and academia to make rapid but informed public health decisions in the face of a new disease.

Mathematical modelling quickly emerged as a strong decision-making tool in the public health arsenal. In recognition of the real-time need for informed modelling guidelines, the Public Health Agency of Canada (PHAC) and the Natural Sciences and Engineering Research Council (NSERC) created the Emerging Infectious Diseases Modelling Initiative (EIDM). The goal was to combine the strengths of Canada's leading experts across a breadth of related fields in order to better respond to the COVID-19 emergency, and to establish a framework that could keep the country prepared for similar situations in the future.

The Mathematics for Public Health (MfPH) network, funded under this initiative, is a pan-Canada partnership between the Fields Institute, the Atlantic Association for Research in Mathematical Sciences, the Centre de Recherches Mathématiques, the Pacific Institute for Mathematical Sciences and many other stakeholders. MfPH builds on the work of the Mathematical Modelling of COVID-19 Task Force funded by the Canadian Institute of Health Research (CIHR) 2019 Novel Coronavirus (COVID-19) rapid research program.

As its top priority, MfPH met to develop novel mathematical tools and analyses to better understand COVID-19 transmission dynamics and support public health decision-making. But it also was committed to boosting future pandemic preparedness by training the next generation of researchers.

Since its establishment in April 2021, Mathematics for Public Health (MfPH) has funded 66 highly qualified personnel (HQP) graduate students and postdoctoral researchers. These HQPs have worked on the multiple research projects of the network, and served as the “frontline workers” of the initiative.

To acknowledge their contributions and to give them an opportunity to build expertise and to establish strong relationships amongst themselves, the MfPH Next Generation was created in September 2021. This group, comprising of early career

researchers, has grown to 80 members and includes modellers, statisticians and epidemiologists interested in using mathematics for public health.

The MfPH Next Generation members are mentored by more senior network members and meet weekly to share knowledge through a biweekly Seminar Series. The MfPH Next Generation Seminar Series provides an inclusive online space for presentations by guest speakers, a journal club and opportunities for members themselves to present their own new work.

This volume is a compendium of their contributions and all chapters are based on presentations given during the MfPH Next Generation Seminar Series. The editors believe this volume should be of interest to public health personnel, applied mathematicians and both current and future generations of researchers at all levels, as it contains state-of-the-art in mathematical modelling and simulations. As well, it should be of interest to graduate students and postdoctoral fellows who are not only interested in theoretical frameworks, but also background introductions, methodologies, interpretations of analyses and applications.

The chapters in this volume cover a variety of mathematical modelling techniques that have been used, or can be used, to support decision-making on public health-related issues, such as resource allocation, impact of climate change on the propagation of communicable diseases, the relationship between human behaviour and disease spread, the projection of disease outbreak trajectories, the evaluation of the efficacy of public health interventions, preparedness and mitigation of emerging and re-emerging infectious disease outbreaks, drug and vaccine development, optimal allocation and distribution of vaccines, and more.

The diseases and public health issues considered in this volume include:

- Mathematical models from the perspectives of mathematical modellers and public health professionals
- Discovering first-principle of behavioural change in disease transmission dynamics by deep learning
- Understanding epidemic multi-wave patterns via machine learning clustering and the epidemic renormalization group
- Contact matrices in compartmental disease transmission models
- Optimal control approaches for public health interventions on an epidemic-viral model in deterministic and stochastic environments
- Modelling airborne disease dynamics: progress and questions
- Modelling mutation-driven emergence of drug-resistance
- Categorical frameworks for modelling with stock and flow diagrams
- Agent-based modelling and its trade-offs: an introduction and examples
- Mathematical assessment of the role of interventions against SARS-CoV-2
- Long-term dynamics of COVID-19 in a multi-strain model

A notable feature of this volume is the inclusion of a variety of model templates and references in a single collection, as well as detailed explanations of the models, methods, algorithms and their assumptions and limitations.

The MfPH Next Generation network benefited enormously from the expertise of the MfPH network members and are grateful for the mentorship and support of MfPH network members in preparing this volume.

The editors would also like to thank all authors and co-authors for their contributions of novel results. All chapters have been peer-reviewed. Many individuals worked hard and collectively behind the scenes.

We would like to specifically thank the following individuals for their help in reviewing chapters and providing technical and professional behind-the-scenes supports: Jesse Knight, Xiaoyan Li, Maritn Grunnill, Idriss Sekkak, Sungju Moon, Gabrielle Brankston, Pengfei Song, Ao Li, Zahra Mohammadi, Sana Jahedi, Shahram Vatani, Arnab Mukherjee, Elisha Are, Woldegebriel Assefa, Bouchra Nasri, Jude Kong, Nicola Bragazzi and Sharmistha Mishra. We would like to also thank the support from the leadership and management team, Sarah Nayani, Deirdre Haskell and Kumar Murty, of the MfPH network and the Fields Institute.

Finally, the editors would like to thank Vignesh Viswanathan, Springer's Project Coordinator, for making the volume development process efficient. And, of course, we especially appreciate our families for their support and love during volume compilation and submission.

Toronto, ON, Canada

Jummy David
Jianhong Wu

Contents

1	Mathematical Models: Perspectives of Mathematical Modelers and Public Health Professionals	1
	Jummy David, Gabrielle Brankston, Idriss Sekkak, Sungju Moon, Xiaoyan Li, Sana Jahedi, Zahra Mohammadi, Ao Li, Martin Grunnill, Pengfei Song, Woldegebriel Assefa, Nicola Bragazzi, and Jianhong Wu	
2	Discovering First Principle of Behavioural Change in Disease Transmission Dynamics by Deep Learning	37
	Pengfei Song, Yanni Xiao, and Jianhong Wu	
3	Understanding Epidemic Multi-wave Patterns via Machine Learning Clustering and the Epidemic Renormalization Group	55
	Shahram Vatani and Giacomo Cacciapaglia	
4	Contact Matrices in Compartmental Disease Transmission Models ..	87
	Jesse Knight and Sharmistha Mishra	
5	An Optimal Control Approach for Public Health Interventions on an Epidemic-Viral Model in Deterministic and Stochastic Environments	111
	Idriss Sekkak and Bouchra R. Nasri	
6	Modeling Airborne Disease Dynamics: Progress and Questions	129
	Arnab Mukherjee, Saptarshi Basu, Shubham Sharma, and Swetaprovo Chaudhuri	
7	Modeling Mutation-Driven Emergence of Drug-Resistance: A Case Study of SARS-CoV-2	161
	Congjie Shi, Thomas N. Vilches, Ao Li, Jianhong Wu, and Seyed M. Moghadas	

8 A Categorical Framework for Modeling with Stock and Flow Diagrams	175
John C. Baez, Xiaoyan Li, Sophie Libkind, Nathaniel D. Osgood, and Eric Redekopp	
9 Agent-Based Modeling and Its Trade-Offs: An Introduction and Examples	209
G. Wade McDonald and Nathaniel D. Osgood	
10 Mathematical Assessment of the Role of Interventions Against SARS-CoV-2	243
Salman Safdar and Abba B. Gumel	
11 Long-Term Dynamics of COVID-19 in a Multi-strain Model	295
Elisha B. Are, Jessica Stockdale, and Caroline Colijn	

Chapter 1

Mathematical Models: Perspectives of Mathematical Modelers and Public Health Professionals



Jummy David, Gabrielle Brankston, Idriss Sekkak, Sungju Moon,
Xiaoyan Li, Sana Jahedi, Zahra Mohammadi, Ao Li, Martin Grunnil,
Pengfei Song, Woldegebriel Assefa, Nicola Bragazzi, and Jianhong Wu

1.1 Natural History of Disease in Humans

The epidemiologic triad of infectious disease is shown in Fig. 1.1b and consists of the interaction between a host (e.g., human), an infectious agent (e.g., virus, bacterium), and the environment in which the opportunity for exposure exists (e.g., contaminated water source) [13, 26]. Diseases can be transmitted via direct (person-to-person contact) or by indirect pathways (e.g., common source or vector), and different pathogens are spread via different routes [12, 13, 26]. For example, many respiratory pathogens can be transmitted by direct contact or via inhalation of

J. David (✉) · M. Grunnil · P. Song · W. Assefa · N. Bragazzi · J. Wu

Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics,
York University, Toronto, ON, Canada

e-mail: jummy30@yorku.ca; grunnill@yorku.ca; song1012@yorku.ca; wassefaw@yorku.ca;
wujh@yorku.ca

G. Brankston

Department of Population Medicine, Ontario Veterinary College, University of Guelph, Guelph,
ON, Canada

e-mail: brankstg@uoguelph.ca

I. Sekkak

Centre de recherche en santé publique and Département de médecine sociale et préventive, École
de santé publique de l'Université de Montréal, Montreal, QC, Canada

e-mail: idriss.sekkak@umontreal.ca

S. Moon

School of Liberal Arts Sciences, and Business, Nevada State College, Henderson, NV, USA

X. Li

Department of Computer Science, University of Saskatchewan, Saskatchewan, SK, Canada

e-mail: xil658@mail.usask.ca

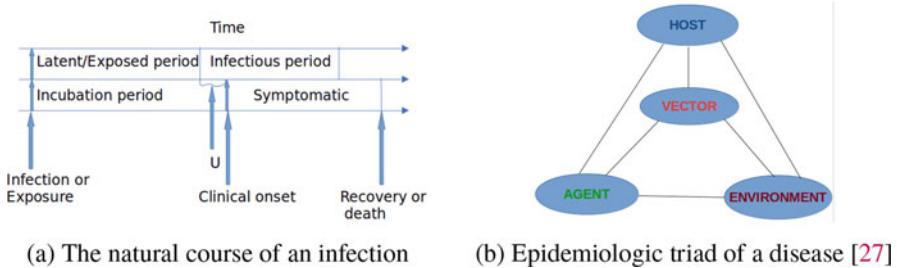


Fig. 1.1 (a) explains the onset of a disease from the infection stage to outcome (removed) stage, while (b) gives the pathways through which diseases are transmitted

aerosols, while other pathogens are spread primarily from infected vectors (e.g., mosquitoes) to humans (e.g., West Nile virus) [13].

Figure 1.1a is a schematic representation of the natural course of an infection. The *latent/exposed period* is the time period during which an individual is infected but shows no signs or symptoms and cannot transmit the pathogen. The *incubation period* is the interval from infection to the time of clinical illness. The *infectious period* is the duration of time during which an individual can transmit a pathogen to others [13, 18, 70]. Mathematical modelers commonly refer to this period as infective period, while public health professionals refer to it as infectious period [3, 13, 18]. The infective and infectious periods will be used interchangeably in this chapter.

Disease characteristics such as the *latent/exposed period*, the *incubation period*, and the *infectious period* are pathogen-dependent and form the basis of infection control measures such as isolation and quarantine which are designed to reduce contact with infectious individuals. When faced with an emerging/re-emerging pathogen, the identification of these characteristics is important in order to design effective control measures to prevent transmission. Pathogens for which transmission is possible prior to symptom onset are particularly challenging to control [3, 26]. Pre-symptomatic transmission is denoted by **U** in Fig. 1.1a. Mathematical models developed for epidemics of emerging/re-emerging pathogens may not account for pre-symptomatic transmission, posing challenges in the development of

S. Jahedi

Department of Biology, McMaster University, Hamilton, ON, Canada

e-mail: jahedis@mcmaster.ca

Z. Mohammadi

Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada

e-mail: zahram@uoguelph.ca

A. Li

Agent-Based Modelling Laboratory, York University, Toronto, ON, Canada

control measures such as quarantine of exposed individuals and isolation of infected cases.

An *epidemic* is defined as the occurrence of disease in a community or region of cases of disease in excess of the level of normal expectancy [13, 18, 26]. Historical examples of epidemics include the 2002–2003 SARS epidemic, the 2014–2016 Ebola virus epidemic in West Africa, and avian flu epidemics [13]. An epidemic that occurs worldwide is referred to as a *pandemic* [26]. *Endemic* refers to the habitual or usual presence of a disease within a given geographic area or population group [13, 18, 26]. The diseases such as malaria, cholera, and typhus are endemic in many parts of the world [13]. Common epidemiological outcomes related to infectious disease are recovery, critical/severe illness, and death as well as epidemic outcomes such as the attack rate and final epidemic size [26].

1.2 Introduction to Mathematical Epidemiology

Mathematical epidemiology is the use of mathematical techniques to understand the spread of diseases in human populations [3]. Mathematical models have been extensively used to study disease transmission dynamics and to extrapolate from epidemiological data in predicting risk. Mathematical modeling in public health allows for an understanding of disease transmission dynamics in order to estimate the rate of epidemic growth, estimate important epidemic parameters, and project the impact of control measures on an epidemic trajectory and its outcomes [13].

One of the most critical concepts in mathematical epidemiology is the basic reproduction number (R_0) which is defined as the average number of secondary infections caused by a typical infectious individual when introduced into a completely susceptible population [3, 13, 74]. R_0 describes the transmissibility of a disease and has a threshold value of 1. A value of more than 1 indicates a growing epidemic, while a value of less than 1 indicates a declining epidemic [3, 13, 74]. When we incorporate interventions aimed at controlling the spread of disease into a model, we use, instead, a *control reproduction number*, denoted as R_c , since control measures decrease reproduction number and therefore decrease the number of secondary infections caused by a single infectious individual.

Reducing R_c below the threshold value of 1 can be achieved by increasing the level of immunity in the population or by introducing interventions to reduce transmission. Thus, the reproduction number can be used to evaluate the effectiveness of public health interventions to reduce disease transmission. For example, the reproduction number can be used to determine the possibility that a disease will be eradicated or eliminated and estimate the impact of public health policies, such as vaccine programs on an epidemic [13]. Models provide a methodical way to estimate R_c , which is important to evaluate the public health measures necessary for disease control and impact on disease transmission [3, 13].

Disease transmission models are developed with a wide range of complexity and the degree of complexity depends on the objective of the model. Models

that make many simplifying assumptions are often designed for analyzing general qualitative model behavior, while detailed and complex models are designed for specific situations and predictive purposes. The use of detailed and complex models for theoretical purposes is limited since they are computationally cumbersome and cannot be solved analytically. While complex models with high strategic value and numerical simulations are needed for detailed planning by public health professionals and policy decision makers, simpler models provide the building blocks of most complex models and may give some useful conclusions [3, 11, 13, 18, 40]. We therefore focus the chapter on simple compartmental models to establish basic concepts.

Public health professionals have contributed immensely to the developments in the modeling of communicable diseases [13]. In 1760, Daniel Bernoulli used life tables and smallpox data to evaluate the impact of smallpox vaccination on life expectancy, resulting in the first known mathematical model [13, 18]. Similarly, public health physicians Sir R.A. Ross, [66], W.H. Hamer [29], A.G. McKendrick, and W.O. Kermack [41–43] laid the foundational concepts of mathematical modeling based on compartmental models in 1900 and 1935, along with J. Brownlee who made important contributions from a statistical perspective. The foundation of disease modeling was based on the concept that individuals can be categorized into mutually exclusive groups by their ability to host and transmit a pathogen.

1.3 Model Formulation and Examples of Some Communicable Disease Models

Different models have been used in different ways to answer public health questions and to assist in the development of public policy. To present a broad overview, we will limit the scope of this chapter to simple compartmental models beginning with epidemic models (with no demographic effects) before expanding on the concepts, incorporating demographic effects to explore the endemic scenario. The Kermack-McKendrick model is a simple compartmental model and makes simplifying assumptions on rates of flow between different classes of individuals in the population. This is the model that will be used to demonstrate modeling concepts. Homogeneous mixing between susceptible individuals and infectious individuals is an assumption of many models; however, contact diary surveys demonstrate that individuals tend to preferentially mix with others in their age group and in specific settings. Examining the network of person-to-person contacts is more realistic for the description of the disease outbreak [3, 28]. The use of *network models* has led to greater improvements in the understanding of epidemic development [3]. Network models are able to show that even if the basic reproduction number is greater than 1, there is a possibility that only a minor outbreak with no full-blown epidemic may occur [3].

An outbreak or epidemic is investigated when three critical variables (the time the exposure took place, the time the disease began, and the incubation period of the disease) are known. One variable can be calculated when other two are known [26]. The study population is divided into different compartments, with assumptions about the nature and time transfer rate from one compartment to another. *Susceptibles* are individuals who have no immunity against the infectious agent and therefore can become infected when exposed. *Infectives* are individuals currently infected and can transmit the infection to susceptible individuals with whom they are in contact. *Removed* individuals are individuals who have immunity against infection and have no effect on transmission dynamics when they are in contact with other individuals [13].

The term *SIR* will be used to describe diseases that confer immunity against reinfection (e.g., recovery from measles or death from plague or rabies). An *SIR* model is one in which the transition of individuals flows from the susceptible class *S* to the infectious class *I* and then to the removed (outcome) class *R*. The term *SIS* will be used to describe diseases with no immunity against reinfection (e.g., common cold) and describes the transition of individuals from the susceptible class *S* to the infectious class *I* and then back to the susceptible class *S* again. Other modeling scenarios, such as *SEIR* (diseases such as tuberculosis) and *SEIS*, include an exposed period between being infected and becoming infectious/infective. *SIRS* models describe diseases with temporary immunity after recovery from infection and *SI* models (diseases such as HIV, herpes) in which there is no recovery from infection [13, 18].

The models presented in this chapter are formulated as differential equations with time t (the independent variable), and transfer rates between compartments are expressed in mathematical terms as derivatives of the sizes of the model compartments with respect to time. It is also possible to generalize to models in which transfer rates depend on the compartment sizes over the past and at the instant of transfer. These will lead to more general types of functional equations (differential-difference equations, integral equations, or integro-differential equations [13]) and will not be considered in this chapter.

1.3.1 Simple SIR Compartmental Models

During the course of an epidemic, there is an initial increase in the number of new infections, which leads to decrease in the number of susceptibles and therefore decreases the number of new infections. This process results in a depletion of susceptible individuals which slows the spread of disease and eventually ends the epidemic [3]. We assume a deterministic epidemic process here and for *SIR* epidemic model, the population under study is divided into three classes *S*, *I*, and *R*. Three papers written by W.O. Kermack and A.G. McKendrick in 1927, 1932, and 1933 proposed simple compartmental models to describe the transmission of communicable diseases. The simple epidemic model that will be considered in this



Fig. 1.2 SIR model schematic [13]

chapter will be a special case of the proposed model by Kermack and McKendrick in 1927 [13, 18], which is given as

$$\begin{aligned}\dot{S} &= -\beta IS, \\ \dot{I} &= \beta IS - \sigma I, \\ \dot{R} &= \sigma I,\end{aligned}\tag{1.1}$$

where β and σ are infection and removal (recovery or death) rate, respectively.

The flowchart in Fig. 1.2 shows the transmission dynamics between compartments. Model (1.1) assumes mass action incidence in which the probability of having a contact is proportional to the size of the population. In mathematical terms, an individual makes contact that is sufficient to transmit infection with βN others per unit time and the total size of the population is assumed to be N . The model focuses on the dynamics of a single epidemic outbreak and therefore assumes a closed population meaning there is no entry into the population (e.g., via births) and that departure exists only through death from the disease (no demographic effects). The model assumes that infectious individuals leave the infective class at rate σI per unit time with recovery rate σ , which gives a mean infectious period of $1/\sigma$. The initial total population $N = S(0)$ and is therefore entirely comprised of susceptible individuals. The probability that an infectious individual made contact with a susceptible individual, resulting in transmission, is given by S/N , and the number of new infections per unit time per infectious individual is calculated as $(\beta N)(S/N) = \beta S$. Thus, the rate of new infections is given by $(\beta N)(S/N)I = \beta SI$, with the transmission rate (per capita) β . Note that for an SIR disease model, the total population is $N = S + I + R$. A similar model implemented without keeping track of the removed individuals is written as

$$\begin{aligned}\dot{S} &= -\beta IS, \\ \dot{I} &= \beta IS - \sigma I;\end{aligned}\tag{1.2}$$

since R does not appear in (1.1), the equation for dR/dt has no effect on the transmission dynamics of S and I [13]. This standard SIR model is presented in many introductory calculus textbooks [13]. The *basic reproduction number* for the model in Eq. (1.2) is calculated as $R_0 = \beta N/\sigma$. Thus, the first infectious individual is expected to infect $\beta N/\sigma$ individuals during their infectious period [13, 74].

Herpes and chronic infections (e.g., HIV) are some of examples of diseases that follow the SI pattern. Similarly, the SIS epidemic model is given as

$$\dot{S} = \delta I - \beta IS;$$

$$\dot{I} = \beta IS - (\alpha + \delta)I, \quad (1.3)$$

where α is the disease-induced mortality and δ is the disease recovery rate with no immunity. Finally, the *SEIR* epidemic model is given as

$$\begin{aligned}\dot{S} &= -\beta IS, \\ \dot{E} &= \beta IS - \nu E, \\ \dot{I} &= \nu E - \sigma I, \\ \dot{R} &= \sigma I,\end{aligned}\quad (1.4)$$

where the exposed individuals leave the exposed class at rate νE per unit time with exposed rate ν , which gives the mean exposed period $1/\nu$.

1.3.2 Simple Endemic Models

The simple endemic model typically uses a longer time scale than an epidemic model and therefore includes demographic processes such as births and deaths. Like the *SIR* epidemic model, the population under study is divided into three classes S , I , and R . Many endemic diseases have caused millions of deaths each year in many parts of the world. For endemic diseases, public health professionals are mostly interested in the number of infectives at a given time, the rate of increase of new infections, potential control measures, and methods to eradicate the disease in a population. The simple endemic *SIR* model is given as

$$\begin{aligned}\dot{S} &= \mu N - \beta IS - \mu S, \\ \dot{I} &= \beta IS - (\alpha + \sigma + \mu)I, \\ \dot{R} &= \sigma I - \mu R.\end{aligned}\quad (1.5)$$

For simplicity, model (1.5) assumes a mass action incidence, similar to the case of epidemic models previously considered. The disease recovery rate is represented by σ and disease-induced mortality is represented by α . Again, for simplicity, we may assume equal birth and death rates as μ and exclude disease-induced mortality (α), such that N is constant. Since $S + I + R = N$, we can determine R if S and I are known, and therefore the model (1.5) can be written as

$$\begin{aligned}\dot{S} &= \mu N - \beta IS - \mu S, \\ \dot{I} &= \beta IS - (\alpha + \sigma + \mu)I.\end{aligned}\quad (1.6)$$

An endemic model that describes diseases with no immunity against reinfection (*SIS* model [3]) is given as

$$\begin{aligned}\dot{S} &= \mu N + \delta I - \beta IS - \mu S, \\ \dot{I} &= \beta IS - (\alpha + \delta + \mu)I,\end{aligned}\tag{1.7}$$

with disease recovery with no immunity represented by δ .

Using model (1.2), we can write the initial exponential growth rate Υ as

$$\Upsilon = \sigma(R_0 - 1).\tag{1.8}$$

Measuring Υ simplifies the estimation of the basic reproduction number (R_0) in Eq. (1.8).

1.3.3 Agent-Based Models

The simple models discussed above treat the population as a whole dividing the population into different compartments. In these *compartmental models*, ordinary differential equations (ODEs) are used to describe the instantaneous flow of individuals between different compartments in the case of an *SIR* epidemic model. In contrast, *agent-based* or, more broadly, *individual-based* models (including *microsimulation* models) adopt the perspective of an individual agent equipped with possible states, transitions between the states, as well as individual actions. See Hunter, MacNamee, and Kelleher [34] for a comprehensive review of different types of agent-based models.

In some sense, agent-based modeling (ABM) can be thought of as Monte Carlo approximations of the solutions to the ODEs in compartmental models. To approximate a compartmental model, agents in each state can be counted in aggregate to form a population compartment. An important advantage of ODE-based models is that they tend to be independent of the population size in view of both the qualitative features of the solutions and required computational costs. On the other hand, because an agent cannot be represented by a fraction in agent-based models, ABM simulation results will approach the ODE solutions only in the limit as the population approaches infinity, and an ABM simulation with a large population size can be quite costly to run computationally. If the modeler's aim is to find closed-form expressions for equilibria or to quickly find the peak in an initial wave of infections, an agent-based model may not be the first choice of tools to consider.

The ABM framework is, however, much more than a mere approximation tool for ODE-based models. The shift in perspective from compartments to individual agents allows for a much greater flexibility in expanding a model to include additional features such as an underlying population network or movement of agents,

introducing heterogeneity in both population and their interactions, embracing stochasticity and exploring emergent behavior as a result, and incorporation of granular and longitudinal data as model inputs.

Suppose a modeler wants to expand an existing *SIR* model by adding a new infectious disease. Without sufficient epidemiological knowledge to eliminate certain interactions possible between the two diseases, a comprehensive model must consider nine compartments, namely, *SS*, *SI*, *SR*, *IS*, *II*, *IR*, *RS*, *RI*, and *RR*, with the first letter of each pair indicating the state with respect to the first disease and the second letter indicating that with respect to the second disease. To add a third disease that also may interact with the other two, the modeler quickly faces the rapid growth of complexity induced by combinatorial explosion. The task is somewhat simpler in the ABM framework. From an individual agent's point of view, the possibility of getting infected with another new infectious disease is just one of the many features the individual can have.

Consider the following analogy. Suppose an essential feature of *personhood* is defined by wearing two pieces of clothing: tops and bottoms. These clothing items are not independent; for example, wearing a short-sleeved shirt may be correlated with wearing shorts. Nonetheless, from an individual's point of view, adding another feature, say, "hats," will not complicate the matter too much. On the other hand, if an entire population has already been compartmentalized based on what kinds of tops and bottoms people wear, adding another feature like "hats" will require a realignment of the existing compartments, complicate the equations, and overall be a great source of headaches.

The flip side of this coin is that, owing to its flexibility, it can be challenging to present and validate agent-based models [32]. Furthermore, as agent-based models tend to evolve in model purpose-specific directions, there may not be enough validating data whose domains are specifically targeted by the models. As mentioned earlier, agent-based models can approximate the kinds of equilibria seen in ODE-based compartmental models. An important design principle to follow is therefore maintaining this upward compatibility with associated compartmental models so that given enough number of agents and forced homogeneity matching results can be obtained for validation. As various different agent-based models built on different platforms are being developed over time, inter-model comparisons and model-group validation against empirical data may play an important role. For example, Eaton et al. [21] conducted a comparison analysis of ten model projections of HIV prevalence in South Africa with the model group consisting of both compartmental and individual-based models.

There are now many software tools specializing in agent-based modeling. One of the earliest programming languages for ABM is NetLogo, which was built in the spirit of the language Logo and likewise uses turtles as agents [80]. Recent uses of NetLogo in infectious disease modeling can be found in [33, 35]. Statistics Canada has long used Modgen [7, 81] for demographic and public health microsimulation and agent-based models including HIVMM (HIV Microsimulation Model) for coevolution of HIV and TB in South Africa, which is being succeeded by a newly developed open-source platform, OpenM++ (openmpp.org). AnyLogic is

a popular general-purpose agent-based modeling software package. Rafferty et al. [64] developed an AnyLogic agent-based model to evaluate the effect of chickenpox vaccination on shingles. Other well-known general-purpose agent-based modeling tools include Repast (repast.github.io) and Swarm (www.swarm.org).

1.3.4 Network Models

Since the pioneering work of Pastor-Satorras and Vespignani [65] established the dynamical mean field equation based on the *SIS* model, the epidemic dynamics in complex networks has been widely investigated by many scholars [5, 52, 53]. Therefore, we consider model (1.5) as a *SIR* epidemic network, where each node describes susceptible (*S*), infected (*I*), or recovered (*R*) compartments with $N = 1$. Given how complicated the network structure gets, the nodes in network are divided into n classes with respect to their degrees, where n denotes the maximum degree of the network. That is to say that the nodes i and j belong to the k -th class if they both have degree k , where $k \in \{1, 2, \dots, n\}$, and we assume that the connectivity of nodes in a network at each time are uncorrelated. Therefore, the dynamical behavior of an epidemic describing the spread on a network is as follows:

$$\begin{aligned}\dot{S}_k &= \mu - \beta k S_k(t) \Theta(t) - \mu S_k(t) \\ \dot{I}_k &= \beta k S_k(t) \Theta(t) - (\mu + \alpha + \sigma) I_k(t) \\ \dot{R}_k &= \sigma I_k(t) - \mu R_k(t),\end{aligned}\tag{1.9}$$

with parameters such as β describing infection of a susceptible individual by an infected neighbor. Recovery of an infected node is denoted by σ and infection-related death is denoted by α . The positive μ denotes the balance between birth rate and natural death rate. The state variables $S_k(t)$, $I_k(t)$, and $R_k(t)$ represent the relative densities of the susceptible, infected, and recovered nodes with degree k , here $k = 1, 2, \dots, n$. $\Theta(t) = \langle k \rangle^{-1} \sum_{k=1}^n kp(k) I_k(t)$ provides the probability that a randomly chosen link originating from a node results in an infected node in a network with degree distribution $p(k)$. Finally, the average degree of the network is denoted by $\langle k \rangle = \sum_{k=1}^n kp(k)$.

1.3.5 Machine Learning Models

Machine learning models aim to estimate, monitor, and predict epidemics using a combination of machine learning algorithms, observed data of infectious diseases, and mathematical models. Machine learning algorithms can be divided into different categories:

1. Supervised learning for labelled observed data and unsupervised learning for unlabelled observed data
2. Classification and regression for supervised learning and clustering for unsupervised learning

However, we will introduce machine learning models by two groups: estimating parameters and estimating hidden states based on the model behavior.

1.3.5.1 Estimating Parameters

A variety of machine learning models have used different algorithms to estimate the parameters in the epidemic model (i.e., the *SIR* model (1.5)) by incorporating the observed data (i.e., epidemiologic surveillance data, social media data, etc.). Commonly used machine learning algorithms include Markov chain Monte Carlo (MCMC) methods and deep learning methods.

Using the *SIR* epidemic model in (1.5) as an example, suppose we have the observed time series of reported cases of a disease (i.e., surveillance data). Machine learning algorithms can be used to estimate a subset of parameters (μ , β , α , and σ) used in the *SIR* epidemic model (1.5) from the observed data. With these estimated parameter values, we can have the trained *SIR* model best fit the observed data. The trained model can then be used to predict forward or to perform counterfactual calculations, such as simulating the impact of interventions.

1.3.5.2 Estimating Hidden States

Another group of machine learning models can estimate hidden states in an endemic model from the observed data. Machine learning algorithms used in estimating hidden states include *filtering methods* (such as Kalman filtering and particle filtering (also called sequential Monte Carlo)) and *deep learning methods of recurrent neural networks* (RNN). For example, in model (1.5), the hidden states are the number of individuals in different compartments, including states *S*, *I*, and *R*. We can estimate the hidden states of endemic models by using the observed data, as well as the trained model to make predictions and counterfactual simulations.

Finally, it is interesting to know that a number of machine learning algorithms can estimate both parameters and hidden states together. For example, the particle Markov chain Monte Carlo (PMCMC) algorithm [2, 23], as a composition of particle filtering algorithm and MCMC algorithm, can efficiently explore high-dimensional parameter space using time series data.

1.4 Qualitative Analysis of Selected Models

1.4.1 Epidemic Model

Model (1.2) with initial conditions $S(0) = S_0$, $I(0) = I_0$, and $S_0 + I_0 = N$ only makes sense when $S(t)$ and $I(t)$ are nonnegative, and then the system ends when either of $S(t)$ or $I(t)$ reaches zero. We notice that $S' < 0$ for all t and $I' > 0$ on the condition that $S > \sigma/\beta$, which then increases I and decreases S for all t . This decrease in S eventually decreases I and I tends to zero. Infective I decreases to zero (no epidemic) whenever $S_0 < \sigma/\beta$, and on the other hand, if $S_0 > \sigma/\beta$, I increases initially to a maximum reached when $S = \sigma/\beta$ and then decreases to zero, which denotes an epidemic. The basic reproduction number for model (1.2) is denoted as $R_0 = \beta S_0/\sigma$, which determines whether an epidemic will occur. The infection dies out whenever $R_0 < 1$, and an epidemic occurs whenever $R_0 > 1$.

Recall that the basic reproduction number is defined as the number of secondary infections caused by the introduction of a single infective into a totally susceptible population of size $N \approx S_0$ during the period of infection of the single infective introduced. In this scenario, βN contacts are made by an infective in unit time, with susceptible individuals producing new infections with a mean infective period $1/\sigma$. The basic reproduction number is then given by $R_0 = \beta N/\sigma$ rather than $\beta S_0/\sigma$. We can also explain this difference by looking at two different ways in which epidemic begins. An epidemic may begin with either a member of a population under study with $I_0 > 0$ and $S_0 + I_0 = N$ or by a visitor from outside of the study population with $S_0 = N$.

The native way to solve a two-dimensional autonomous system of differential equations like model (1.2) is to find the equilibria and determine stability by linearizing about each equilibrium. Nevertheless, model (1.2) has a line of equilibria (i.e., every point with $I = 0$ is an equilibrium), and it is impossible to use this method since the linearization matrix produces a zero eigenvalue at each equilibrium. We therefore use a different method to analyze the system (1.2). The sum of equations S and I in (1.2) gives

$$(S + I)' = -\sigma I.$$

We can see that $(S + I)$ decreases to a limit, and since $(S + I)$ is a nonnegative smooth function, we could show that its derivative approaches zero, from which it can be concluded that

$$I_\infty = \lim_{t \rightarrow \infty} I(t) = 0.$$

Integrate the sum of the two equations of (1.2) from 0 to ∞ to have

$$\sigma \int_0^\infty I(t) dt = S_0 + I_0 - S_\infty = N - S_\infty,$$

$$\int_0^\infty I(t)dt = \frac{N - S_\infty}{\sigma}, \quad (1.10)$$

which implies that $\int_0^\infty I(t)dt < \infty$.

Divide the first equation of (1.2) by S and integrate from 0 to ∞ to have

$$\log \frac{S_0}{S_\infty} = \beta \int_0^\infty I(t)dt,$$

and by substituting Eq. (1.10), we have

$$\log \frac{S_0}{S_\infty} = \beta \frac{N - S_\infty}{\sigma} = \frac{\beta N}{\sigma} \left[1 - \frac{S_\infty}{N} \right] = R_0 \left[1 - \frac{S_\infty}{N} \right]. \quad (1.11)$$

Equation (1.11) is known as the *final size relation*. It gives an estimate of the total number of infections over the course of the epidemic from the parameter in the model [10, 12] and shows the relationship between the basic reproduction number and the size of the epidemic. The final size of the epidemic ($N - S_\infty$) is always described in terms of the attack rate/ratio ($1 - S_\infty/N$). We can generalize the final size relation (1.11) to the epidemic model with more complex compartment than the simple *SIRmodel* (1.2), including model (1.4) with exposed periods, models with treatment, models involving quarantine of suspected individuals, and isolation of diagnosed cases. For example, an epidemic with proportion of susceptibles $S_0 = 0.999$, and $S_\infty = 0.35$ as in Fig. 1.3a and substituting into Eq. (1.11), gives the estimate $\beta/\alpha = 1.61$ and $R_0 = 1.61$.

1.4.2 Endemic Model

We can determine a disease-free equilibrium (DFE) of the endemic model (1.6) by setting $\dot{S} = \dot{I} = 0$:

$$\begin{aligned} \mu N - \beta IS - \mu S &= 0 \\ \beta IS - (\alpha + \sigma + \mu)I &= 0. \end{aligned} \quad (1.12)$$

We therefore have the disease-free equilibrium (DFE) as $(S, I) = (N, 0)$, and the endemic equilibrium point (EEP) as $(S, I) = \left(\frac{(\alpha + \sigma + \mu)}{\beta}, \frac{\mu(\beta N - (\alpha + \sigma + \mu))}{\beta(\alpha + \sigma + \mu)} \right)$, which exists only when $(\alpha + \sigma + \mu) < \beta N$.

We can analyze the stability of the above equilibria by the theorem below:

Theorem 1 Let the basic reproduction number be $R_0 = \frac{\beta N}{\alpha + \sigma + \mu}$, then $R_0 < 1$ shows that EEP does not exist, and for all positive initial conditions, we have

$\lim_{t \rightarrow \infty} (S(t), I(t)) = (N, 0)$, and the disease dies out. Also, if $R_0 > 1$, then for all positive initial conditions,

$$\lim_{t \rightarrow \infty} (S(t), I(t)) = \left(\frac{(\alpha + \sigma + \mu)}{\beta}, \frac{\mu(\beta N - (\alpha + \sigma + \mu))}{\beta(\alpha + \sigma + \mu)} \right) = \left(\frac{1}{R_0} N, \frac{\mu}{\beta}(R_0 - 1) \right),$$

and the disease persists in the population.

We can interpret the basic reproduction number $R_0 = \frac{\beta N}{\alpha + \sigma + \mu}$ (the average number of cases produced when a case is introduced into a totally susceptible population) as the product of:

- β , the probability of contracting the disease when a potentially infecting contact occurs
- $\frac{1}{\alpha + \sigma + \mu}$, the mean time spent in the infectious class when subject to the competing risks of natural death, recovery, and disease-induced death

1.4.3 Network Model

Let $\Gamma = \{(S_1, I_1, Q_1, \dots, S_n, I_n, Q_n) \in R_+^{3n} | S_k + I_k + Q_k = 1, k = 1, 2, \dots, n\}$. The stability analysis is performed in Γ , which is a positive invariant set guaranteed by the following result:

Obviously, the disease-free equilibrium of model (1.9) is $E^0 = (1, 0, 0, 1, 0, 0, \dots, 1, 0, 0) \in \mathbb{R}^{3n}$. Letting the right hand of model (1.9) be zero, we see that its endemic equilibrium $E^* = (S_1^*, I_1^*, Q_1^*, \dots, S_n^*, I_n^*, Q_n^*)$ is determined by the following algebraic equations:

$$S_k^* = \frac{\mu + \alpha + \sigma}{\beta k \Theta^*(t)} I_k^*, \quad I_k^* = \frac{\mu \beta k \Theta^*(t)}{\mu(\mu + \alpha + \sigma) + (\mu + \sigma) \beta k \Theta^*(t)}, \\ R_k^* = \frac{\sigma}{\mu} I_k^*.$$

Therefore, we have the following self-consistent equation with respect to

$$\Theta^* = \langle k \rangle^{-1} \sum_{k=1}^n k p(k) I_k^* \\ = \frac{1}{\langle k \rangle} \sum_{k=1}^n \frac{\mu \beta k \Theta^*}{\mu(\mu + \alpha + \sigma) + (\mu + \sigma) \beta k \Theta^*(t)} \quad (1.13) \\ = \Psi(\Theta^*) \quad (1.14)$$

It is easy to check that $\Psi(0) = 0$, $\Psi(1) < 1$. Therefore, a nonzero $\Theta^* \in (0, 1)$ exists if $\Psi(0) > 1$, which yields to

$$\frac{\beta}{\mu + \alpha + \sigma} \frac{\langle k \rangle^2}{\langle k \rangle}$$

We can define the basic reproduction number as follows:

$$R_0 = \frac{\beta}{\mu + \alpha + \sigma} \frac{\langle k \rangle^2}{\langle k \rangle} \quad (1.15)$$

Lemma 1 *The set Γ is positive invariant for model.*

The basic reproduction number R_0 , which is an important measure in the investigation of an epidemic model and its stability, entirely dominates the dynamical behavior of model (1.9), as shown by the following two theorems.

Theorem 2 *If $R_0 < 1$, then the disease-free equilibrium E_0 of model (1.9) is globally asymptotically stable, i.e., the infection will gradually die out.*

Theorem 3 *If $R_0 > 1$, then the endemic equilibrium E^* of model (1.9) is globally asymptotically stable, i.e., the infection spreads and becomes endemic.*

1.5 Quantitative Analysis

The *SIR* model, which is one of the most basic of all epidemiological models, depends on calculating the percentage of the population in each of the susceptible, infected, and removed/recovered classes and determining the transmission rates between them. Consider the simplest form of an epidemic (ignoring births and deaths) as in Eq. (1.1) in which there are only two transitions: infection (individuals progress from susceptible to the infected class) and recovery (individuals progress from infected to the recovered class). For simplicity, it is often assumed that individuals infected with a disease recover at a constant rate [41], whereas it is generally assumed from epidemic data that the *per capita* rate of a given susceptible individual becoming infected is proportional to the prevalence of the infection in the population [26]. The simple model in (1.1) requires modelers to estimate two parameters (the infection transmission rate β and recovery rate σ) demonstrating the basic relationship between models and statistics. The usefulness of a model depends on good epidemiological data to inform the estimation of parameters.

Once the two parameters have been estimated, the *SIR* model can project an epidemic which follows the pattern in Fig. 1.3a: the number of cases (red) initially increases until the number of susceptible individuals (blue) has been adequately depleted. This process continues until the number of infected individuals eventually decreases and the number of removed individuals increases (green), leading to extinction of the epidemic. The numerical simulations of *SIR* model (1.1) shown in

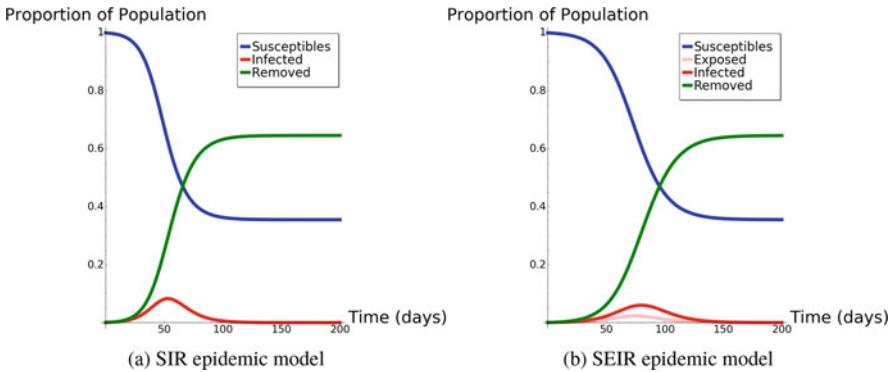


Fig. 1.3 Results of numerical solutions of the *SIR* (a) and *SEIR* (b) epidemic model which predict the rate of change of susceptible, exposed, infected, and removed over time and compare quantitative behaviors of the two models. The simulations show basically the effect of exposed period on the behavior of the model

Fig. 1.3a produce three general predictions that are of importance to public health and have policy implications. Predictions from this simple model are supported by many more complicated models with numerous parameters [1, 40]. For example, if Fig. 1.3a assumes the numerical simulation for the total proportion of the population $N = 1$, with $S_0 = 0.999$, $I_0 = 0.001$ and with $\beta = 0.3$, $\sigma = 0.187$. We can therefore predict the following:

1. The basic reproduction number $R_0 = \beta/\sigma$, which is the parameter that controls the epidemic dynamic. The value of $R_0 < 1$ denotes an epidemic that is in decline due to the inability of the epidemic to sustain the transmission dynamics, whereas $R_0 > 1$ indicates the possibility of an epidemic. For the example shown in Fig. 1.3a, R_0 is estimated to be approximately 1.6 which is dependent on both the population and infection.
2. In general, the proportion of the population that remains susceptible at the end of epidemic becomes very small for large values of R_0 , but for the scenario of $R_0 = 1.6$, approximately 62% of the population is expected to be infected during an epidemic. A more complicated model with many parameters may change the precise value of the proportion infected, but the general idea continues to hold.
3. Susceptibility could be reduced through vaccination, thereby decreasing the spread of infection in the population. An epidemic could be prevented by vaccinating some proportion of the population, and endemic infection could be eradicated or pandemic prevented if a proportion $1 - 1/R_0$ of the population is successfully immunized [1]. In the current example, we would need to immunize approximately 38% of the population to stop the endemic. This value can be reduced if vaccination is sensibly targeted with a more complicated model [1, 40].

We can expand the epidemic scenario to include an exposed class as in Eq. (1.4). In this case, there are three transitions: exposure (individuals progress from susceptible to the exposed class), infection (individuals progress from exposed to the infected class), and recovery (individuals progress from infected to the recovered class). Initiating the model in (1.4) requires modelers to estimate three parameters (the infection transmission rate β , exposed rate ν , and recovery rate σ).

Once the three parameters have been estimated, the *SEIR* model can project an epidemic which follows the pattern in Fig. 1.3b: the number of cases (red) initially increases but lower than in *SIR* model from Fig. 1.3a. The epidemic was sustained for approximately 100 and 130 days in the *SIR* and *SEIR* model, respectively (Fig. 1.3a and b). The numerical simulations of the *SEIR* model (1.4) shown in Fig. 1.3b produce three general predictions like in the *SIR* model (1.1).

From the previous example, if Fig. 1.3b assumes the numerical simulation for total proportion of population $N = 1$, with $S_0 = 0.999$, $E_0 = 0$, $I_0 = 0.001$ and with $\beta = 0.3$, $\sigma = 0.187$, $\nu = 0.5$, the basic reproduction number $R_0 = \beta/\sigma$ is similarly estimated to be 1.6. The proportion of the susceptible population at the end of epidemic in both models is not significantly different as the same number of population would need to be vaccinated (NNT= 38%) to prevent one less case.

In our example, both the *SIR* and *SEIR* models gave similar results and therefore the *SIR* model achieves the same outcome with fewer parameters. We can improve the predictive value of a model while increasing the number of estimated parameters required, by considering more complicated models which incorporate heterogeneous mixing and the potential for overdispersion of cases (i.e., superspreading) [11], metapopulation studies [4], age of infection [10], residence time [8, 11], and mixing patterns through network models [63].

1.6 Review of Mathematical Models of Selected Communicable Diseases and Their Impacts on Policy- and Decision-Making

Much of the scientific literature provides convincing evidence of the impact of mathematical and simulation models on disease transmission and in public health. Mathematical models have been extensively used for estimating the possible effect of intervention strategies and advising policy decision-making [70]. Infectious disease outbreaks in the past have shown the necessity for planning and epidemic readiness to ensure health resources are optimally distributed and for decreasing illness and death among target populations [70]. The following sections will describe mathematical modeling of known infectious diseases and their impact on policy- and decision-making.

1.6.1 SARS 2003 Pandemic Models

Soon after SARS has been identified by the World Health Organization, several simple compartmental models at population level were proposed with varying success. The *SIR* model by Choi and Pak [17] for an early-stage epidemic introduced simplifications such as homogeneous population mixing and zero intervention measures. To simulate the transmission dynamics in Beijing where the virus was originally found, Wang and Ruan [76] proposed a model consisting of six subpopulations: susceptible, exposed, quarantined, suspect infectives, probable infectives, and recovered. The model was simplified to a two-compartment suspect-probable model and a single-compartment probable model to obtain estimates using limited data.

Additional, more complex factors were gradually added to such compartmental models. Zhou et al. [85] incorporated diagnosis and quarantine into the model as intervention measures to control the spread of disease. Webb et al. [77] distinguished individuals in the hospital setting including health-care workers and patients from the general public due to the different transmission dynamics within hospitals. Simulation results indicate that the key to containing SARS was a combination of moderate quarantine and strict hospital infection control procedures. While these studies focused on the asymptotic behavior of deterministic ODE-based models, evidence of apparent stochasticity of disease transmission can be found in the contrasting experiences of Vancouver, BC, and Toronto, ON. These two cities with similar infrastructure and response strategies exhibited vast differences in the disease containment outcomes [14]. Aya, Aldila, and Handari [6] studied a model equipped with stochastic differential equations with perturbation parameters to represent the effect of random factors in infection probability and recovery rates on disease spread. This model also allowed reinfection and intervention measures such as mask-wearing and medical treatments. While the model itself was unable to accurately capture daily incidence rates, the study was able to conclude that the stochasticity of infection probability had greater effects on the outcomes than that of recovery rates.

One important characteristic of SARS-CoV outbreaks identified from early studies is the occurrence of superspreading events (SSEs). Li et al. [48] adopted probability distributions of the incubation period as well as time from onset of symptoms to hospital admission to simulate intermittent peaks what would suggest the occurrence of SSEs. In recent years, the adaptation of agent-based modeling allowed for a more granular approach to modeling SSEs (see [44] for agent-based modeling of MERS-CoV).

Another notable development in the study of SARS was the recognition of the significant role of airborne transmission in SARS outbreaks [67], which was initially dismissed as a significant contributing factor [14]. Yu et al. [84] found that the predictions based on computational fluid dynamic (CFD) simulations of airflow corresponded well with the distribution of exposure risk in different buildings within the Amoy Gardens residential complex, leading to the airborne hypothesis

of SARS. Side-by-side comparisons of meteorological data and outbreak data also suggested that a reduction in environmental temperature might have led to heightened survival of aerosolized virus particles and impeded their vertical dispersion [83]. Combined with the northeasterly winds which transported the aerosolized virus to other residential blocks, these environmental conditions likely led to the notorious cluster event at Amoy Gardens [83]. A similar CFD simulation of aerosol dispersion within the patient wards of the Prince of Wales Hospital suggested that airborne transmissions played a crucial role in hospital outbreaks [49]. The airborne hypothesis was further substantiated by the air sampling data from SARS units of four Toronto health-care facilities at the height of the Toronto outbreak [9], highlighting the importance of ventilation among different intervention measures.

Armed with the knowledge about the airborne nature of SARS, Naheed, Singh, and Lucy [60] developed a hybrid compartmental population model with one-dimensional spatial diffusion (i.e., ventilation) components. The model combines the five-class compartmental model of Chowell et al. [19] consisting of susceptible, exposed, infected, diagnosed, and recovered populations with spatiotemporal diffusivity terms. Their numerical results suggest that the inclusion of diffusion changes distributions of susceptible, exposed, and infected populations leading to more individuals getting infected in a shorter period of time and quickly reaching a maximum. This study also supports ventilation and air filtration as a practical measure for reducing the intensity of a disease outbreak.

At a more granular level, Lei et al. [46] employed a mathematical model to simulate the spread of viral particles among individuals on board an airplane. The model consists of droplet equations resembling microphysics equations and considers different virus concentrations following exposures via airborne routes, via close contact routes, and via fomites, taking into account the known viral properties of SARS, norovirus, and H1N1. It was found that while transmission via fomites played the most important role followed by close contact, airborne routes were still a significant factor for SARS compared to norovirus and H1N1, therefore requiring different prevention strategies compared to other viral pathogens.

The respiratory illness SARS is transmitted through contact from infected people to others. Traditional compartmental models assume the population groups are fully mixed, meaning that each individual has an equal chance of spreading the disease. However, contact patterns are highly heterogeneous in reality. Meyers et al. [55] introduced network theory into SARS transmission model where networks were formed by physical contacts among people. This work provides insight into the diversity of outbreaks around the world.

1.6.2 Pandemic Influenza Models

Influenza is a respiratory virus that mainly infects the nose, throat, bronchi, and, on rare occasions, the lungs. The symptoms of infection typically last for 1 week and include fever, fatigue, headache, cough, and sore throat. Seasonal influenza

epidemics occur annually and are typically caused by influenza A or B for which there is some prior immunity in the population due to previous exposures or vaccinations. Pandemic influenza occurs when a new influenza virus A subtype for which there is no immunity in the population is introduced and spreads worldwide. Examples of severe influenza outbreaks [62] are discussed. Influenza subtype H3N8 is thought to have caused a pandemic in 1889–1890 with an estimated case fatality ratio of 0.1–0.28% and a basic reproduction number of 2.1 (IQR, 1.9–2.4) [73]. One of the most familiar influenza pandemics that occurred in 1918–1920 was caused by H1N1 and resulted in a case fatality ratio greater than 2.5% [72] and an estimated basic reproduction number of 2.0 (IQR 1.7–2.3) [56]. Another well-known pandemic was the Asian flu caused by H2N2 in 1957–1958, which resulted in an estimated case fatality ratio of 0.2% and a basic reproduction number of 1.5 [82].

The dynamics of influenza virus transmission using mathematical models, particularly compartmental frameworks, have been well-described in the scientific literature. In [69], a mathematical model was developed for predicting the global spread of influenza, using the forecast of the 1968–1969 “Hong Kong” influenza pandemic among 52 major cities worldwide. The effect of seasonality on the dynamics of influenza transmission was examined by introducing a fourth class (C) for cross-immune individuals into a SIR epidemic model, where they investigated a seasonality effect using bifurcation analysis [15]. This approach was demonstrated to be capable of predicting a wide range of complex temporal patterns that are typical for influenza. Control theory and compartmental modeling have been used to examine the impact of vaccination on the number of asymptomatic infections and the final epidemic size of an influenza pandemic [36]. In order to explore multiple intervention measures, authors [75] developed an influenza model with imperfect vaccination, media coverage, and antiviral therapy to determine the most effective strategy and combinations of methods to prevent influenza transmission. In addition, several authors have investigated the coexistence of two different infectious diseases including influenza, where Wilasang et al. [79] investigated the competitive evolution of two different subtypes of influenza viruses H1N1 and H3N2. Meanwhile, authors in [61], using stability analysis and control theory, studied the effects of competitive outcomes between and within hosts on the dynamics of a SARS-CoV-2 and influenza co-infection.

1.6.3 SARS-CoV-2 Pandemic Models

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a strain of coronavirus that causes COVID-19 (coronavirus disease 2019), the respiratory illness responsible for the ongoing COVID-19 pandemic. SARS-CoV-2 was first identified in the city of Wuhan, Hubei, China; the World Health Organization declared the outbreak a public health emergency of international concern on 30 January 2020 and a pandemic on 11 March 2020.

At the beginning of COVID-19 epidemic, SARS-CoV-2 models focused on forecasting and evaluating the disease severity by estimating incubation period, the basic reproduction number, and other epidemiological parameters [47, 71]. As time went on, many mathematical models were developed to investigate the impact of non-pharmaceutical interventions on disease transmission. Examples of these models include the following:

- Lockdown: Authors explored a local strategy of reopening (and re-closing, as needed) schools and workplaces county by county, based on triggers for county-specific infection prevalence, compared with a global strategy of province-wide reopening and re-closing, according to triggers for province-wide infection prevalence. Model results demonstrated that local strategies outperform global strategies in the early epidemic stage but only if testing rates are high and the trigger prevalence is low [39].
- Isolation: A stochastic transmission model was developed to explore the feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts [31].
- Testing and social distancing: A model of individual-level transmission stratified by setting (household, work, school, or others) based on BBC Pandemic data from 40,162 UK participants was developed to investigate the effect of a range of different testing, isolation, contact tracing, and physical distancing scenarios [45]. The authors examined the impact on the effective reproduction number and the number of contacts that would be newly quarantined each day under different strategies.
- Mobility restriction: A global metapopulation disease transmission model was used to project the impact of travel limitations on national and international transmission of SARS-CoV-2, based on internationally reported cases [16]. They found that travel quarantine in Wuhan delayed the overall epidemic progression by only 3 to 5 days in Mainland China but had a more pronounced effect at the international scale, where case importations were reduced by nearly 80% until mid-February.
- Human behavior changes: An investigation of the impact of human behavior changes on transmission of SARS-CoV-2 demonstrated that awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations [78].
- Health resources: *SEIAR* and *SEIA – CQFH* warehouse models were developed to simulate the two-period epidemic in Wuhan and to quantitatively assess the effectiveness of joint measures led by Fangcang shelter hospitals in response to the COVID-19 epidemic in Wuhan, China [37].
- Reopening: [24] consider epidemiological modeling for the design of COVID-19 interventions in university populations and proposed an epidemiological model that formed the basis for Cornell University's decision to reopen for in-person instruction in fall 2020.
- Ranking of NPIs: Investigators assessed the effectiveness of non-pharmaceutical interventions (NPIs) to mitigate the spread of SARS-CoV-2 [30]. They quantified the impact of 6068 hierarchically coded NPIs implemented in 79 territories on

the effective reproduction number, R_t , of COVID-19 and proposed a modeling approach that combines four computational techniques merging statistical, inference, and artificial intelligence tools.

Once COVID-19 vaccines became available, many researchers used mathematical models to explore optimal vaccination strategies [20, 25] and explored the indirect effect of vaccination on populations.

1.6.4 HIV Models

In spite of the control and prevention tools for HIV, both developed and developing countries continue to experience an increase in the incidence and prevalence of HIV [50, 58]. Proper use of highly active antiretroviral therapy (HAART) has reduced the morbidity and mortality of HIV-infected individuals [50, 57, 58], and the increasing use of HAART has shown an obvious positive impact in efficacy and effectiveness [57]. HIV treatment as prevention (TasP) [57] has also been used to prevent or control progression to AIDS, prevent premature death among HIV-infected individuals, and therefore reduce HIV transmission [58].

Lima et al. developed a *semi-deterministic* dynamic transmission mathematical model which incorporates multiple sources of infection to assess the impact of expanding access to HAART (guidelines, coverage rates, HAART coverage, and adherence) on new HIV infection and on associated costs in British Columbia, Canada [50]. GAMMs (an advanced statistical method) was used to allow for change in transmission probabilities over time. The model projected that the existing level of coverage (50%) and 78.5% adherence would result in an increase in the number of newly identified HIV infections from 421 to 462 between 2006 and 2030. An increase in HAART coverage to 75%, 90%, and 100% of those in medical need would result in a reduction in the annual number of newly identified HIV infections. Furthermore, a decrease in level of adherence to below 40% (in the case of 50% coverage) would expectedly increase the number of individuals testing newly positive for HIV by 10% per year. Increasing adherence to levels above 80% in the same situation, the number of individuals testing newly positive for HIV was projected to decrease by 0.3% per year. The results suggest a reduction in the growth of the HIV epidemic and associated costs by changing policy to improve access to HAART programs.

Granich et al. used both *stochastic* and *deterministic* transmission models to investigate the effect of testing all people aged 15 years and older, for HIV annually and with immediate placement of positive individuals on antiretroviral therapy (ART) [27]. Data from South Africa was used to explore epidemic scenarios, with an assumption of a heterosexual HIV transmission. A stochastic model was used to investigate the impact of different treatment plans and model parameters on R_0 , while the deterministic model was used to investigate the impact of different HIV testing and treatment methods on the overall dynamics of the HIV epidemic. The

study estimated an acceleration in the change from the current endemic stage (most HIV-infected adults not on ART) to the elimination stage (most HIV-infected adults on ART), within 5 years. The results suggested that universal voluntary HIV testing with immediate ART could decrease HIV incidence and mortality to < 1 case per 1000 people per year by 2016 in 10 years of complete discharge of the program and therefore decrease the prevalence of HIV to < 1% in 50 years. They estimated that in 2032, the cost of the current and the theoretical plan would both be US \$1.7 billion; and after this time, there will be a decrease in the theoretical plan and an increase in the current plan.

Their model predicted a major effect on acute generalized HIV/AIDS epidemics with universal voluntary testing and immediate ART incorporated with current prevention strategies. The major limitation in this study is the data used which do not account for some important variables. There is need for better data for appropriateness and use of universal voluntary HIV testing, the contagiousness of people on ART, adherence, change in behavior after beginning ART, and emergence of resistance rates. There is also need for comprehensive economic analysis of the suggested program to explore the economic impacts of the theoretical plan. In addition, they suggested that there is need for stakeholders to do a proper evaluation and consultation of the theoretical approach, and the role of ART.

Eaton et al. [22] compared 12 independent mathematical models from researchers who had previously developed mathematical models, to assess the impact of antiretroviral therapy (ART) on HIV incidence in South Africa. They concentrated on the need for understanding the effect of ART on HIV transmission, in addition to the generally known benefits of reduction in morbidity and mortality among patients on treatment. Four models were agent-based *microsimulation*, and the other eight were *deterministic* compartmental models stratified by the characteristics of individual and HIV infection status. This array of mathematical models was used to understand the level of agreement of different mathematical models on the epidemiological effect of expanded access to ART and inform public policy about HIV treatment as prevention (TasP) in South Africa. Results of the 12 models were uniform for reduction in HIV incidence in the short-term prediction (8 years) but varied in the long-term prediction.

In addition, mathematical modeling was used by Nah et al. [59] to examine the test-and-treat policy for controlling HIV/AIDS. This modeling study showed a decrease in HIV incidence with definite coverages of diagnosis, care, and continuous treatment, and no obvious decrease in the HIV prevalence was found, but there was an increase in long-term cost of ART. The Joint United Nations Programme on HIV/AIDS (UNAIDS) initiated the notion of HIV treatment cascade to know and fill breaches in the continuum of services for testing, care, and effective treatment. According to the 21st International AIDS Conference in Durban, South Africa, the UNAIDS news led to a world slogan of “90-90-90” target by 2020. This target aims that 90% of HIV-infected individuals are aware of their HIV status, 90% of individuals who know their status are on treatment, and 90% of individuals on treatment experience a suppressed viral load [38]. UNAIDS aims to attain individual target of 95-95-95 by the year 2030.

1.6.5 HCV Models

Treatment for hepatitis C virus (HCV) has been available for more than a decade; however, the length of treatment, issues of tolerability, and pill burden are barriers for accessing treatment for the most susceptible population [68]. The following is a description of a selection of mathematical models of HCV and their pertinence to the formulation of public health policy.

Rozada et al. [68] developed a *deterministic* compartmental model of HCV with 35 ordinary differential equations (ODEs) among people who inject drugs (PWID) with a detailed model of fibrosis progression to inform policy in treating individuals based on fibrosis level. Their work analyzed the stability properties of the solutions to the model, to understand the conditions in which HCV could be eliminated among PWID population of British Columbia, Canada. The model took the practical characteristics of HCV epidemic (such as reinfection, reduction of risk after being successful in eliminating the disease, time to diagnosis) into consideration. The integration data were linearly interpolated and evaluated at annual intervals to obtain data outputs. Powell hybrid optimization algorithm was used to fit the force of infection and mortality rate parameters to known prevalence and incidence rates at the endemic steady-state equilibrium. The parameter space of treatment and testing rates that would eliminate HCV ($R_0 < 1$) were determined by the new treatment. Sensitivity analysis was done and the contact rate was found to be the most sensitive parameter, second to the period of the acute phase.

From the perspective of public health professionals, the model results showed that it is possible to eliminate HCV when more effort is put towards reducing contact rate, for example, through harm reduction. One of the limitations of the model is the inability to stratify according to the level of risk behavior, as there could be differences in risk behavior among high- and low-risk PWID. For this reason, the model may not be generalizable since the data from the PWID cohorts in BC, Canada, were used. In this region, the HCV prevalence is around 65%, and results from this model may not be the same for a region where HCV prevalence is relatively higher or lower than 65%. The authors advised public policy-makers to increase treatment coverage and decrease the time to access treatment so as to weaken the endemic solution. Lima et al. in [51] expanded the model in [68] to study the effect of the new therapies on incidence, prevalence, and mortality.

Martin et al. [54] predicted the possible effect of treatment scale-up in the age of direct-acting antiviral (DAA) treatments on people who inject drugs using a parameterized *deterministic* HCV transmission and treatment model. Three chronic HCV prevalence environments (Edinburgh, UK (25%), Melbourne, Australia (50%), Vancouver, Canada (65%)) were considered, and the treatment rate to decrease the prevalence of chronic HCV by half or three-quarters within 15 years was predicted. The population in the model was stratified according to the state of infection. Uncertainty and sensitivity analyses were done to explore the impact of individual model assumptions in the parameters. The initial treatment of HCV was projected to reduce the prevalence of HCV in Melbourne and Vancouver by < 2% and in Edinburgh by 26% within 15 years. Scale-up of treatments of Edinburgh, Melbourne, or Vancouver, respectively, to 22, 54, or 98 per 1000 PWID annually

was predicted to decrease the prevalence by three-quarters in 15 years. Model predictions demonstrated that scaling up of treatment to 20 per 1000 PWID yearly would reduce the prevalence in 15 years by 69% (95% CrI, 54%–83%) and 23% (95% CrI, 17%–32%) in Edinburgh and Melbourne, respectively, but only 9% (95% CrI, 7%–15%) in Vancouver.

The authors suggested a need for higher treatment rates (> 40 per 1000 PWID per year) to reduce the prevalence by > 20% in Vancouver within 15 years. Their work also suggested interferon-free DAA HCV treatment as prevention, to be a possible option to decrease the future problem of HCV-associated diseases that are of public health interest in a situation where HCV prevention programs are unavailable.

1.7 Model Algorithms for a Simple SIR Model

1.7.1 Python Code

```

1 from numpy import zeros, linspace
2 import matplotlib.pyplot as plt
3 # Time unit: 1 Day
4 mu=7.8/(1000*365)
5 N=0
6 beta = 10./(40)
7 alpha = 3./(15)
8 sigma = 0.1
9 dt = 0.1
10 D = 30           # Simulate for 30 days
11 N_t = int(D/dt)  # Corresponding no of days
12
13 t = linspace(0, N_t*dt, N_t+1)
14 S = zeros(N_t+1)
15 I = zeros(N_t+1)
16 R = zeros(N_t+1)
17
18 # Initial condition
19 S[0] = 0.5
20 I[0] = 0.3
21 R[0] = 0.2
22 #Basic reproduction number
23 R0= beta / (mu + alpha + sigma)
24 # Step equations forward in time
25 for n in range(N_t):
26     S[n+1] = S[n] + dt*(mu*N-beta*S[n]*I[n])
27     I[n+1] = I[n] + dt*(beta*S[n]*I[n] - (alpha+mu+sigma)*I[n])
28     R[n+1] = R[n] + dt*(sigma*I[n]-mu*R[n])
29
30 fig = plt.figure()
31 l1, l2, l3 = plt.plot(t, S, t, I, t, R)
32 fig.legend((l1, l2, l3), ('S', 'I', 'R'), 'upper left')
33 plt.xlabel('Days')
34 plt.show()
```

Listing 1.1 Python example

1.7.2 Julia Code

In this section, we will run a deterministic simulation of the SIR model. Transmission dynamics can be divided into two parts: latent dynamics and observation dynamics. Observation dynamics are concerned with data, such as accumulated cases and death cases.

Latent dynamics are given as in the equations;

$$\begin{cases} \frac{dS}{dt} = -\beta SI, \\ \frac{dI}{dt} = \beta SI - \gamma I, \\ \frac{dR}{dt} = \gamma I, \end{cases}$$

and observation dynamics (e.g., accumulated cases also given as)

$$\frac{dH}{dt} = \beta SI,$$

or sometimes also defined as

$$\frac{dH}{dt} = \gamma I,$$

or with delay

$$\frac{dH}{dt} = \gamma I(t - \tau),$$

Julia codes are as follows:

```

1 using DifferentialEquations
2 using Plots
3 function SIR(du, u, p, t)
4     beta, gamma= p
5     S, I, R, H = u
6     du[1] = - beta * S * I
7     du[2] = beta * S * I - gamma * I
8     du[3] = gamma * I
9     du[4] = beta * S * I
10 end

```

(continued)

```

11 tmax = 20.0
12 dt = 0.1
13 tspan = (0.0,tmax)
14 u0 = [10,0.1,0.0,0.0]
15 p = [0.2,0.1]
16 prob_ode = ODEProblem(SIR,u0,tspan,p)
17 sol = solve(prob_ode,Tsit5(),saveat=dt);
18 plot(sol, label = ["S" "I" "R" "H"], title = "Deterministic
    Simulation of SIR Model")

```

Listing 1.2 Julia example

1.7.3 R Code

```

1 #1. Call libraries
2 require(deSolve)
3 library(ggplot2)
4 #2.SIR model
5 SIR <- function(time, x, params){
6   S <- x[1] # susceptible
7   I <- x[2] # infected. Contagious.
8   R <- x[3] # recovered
9   with(as.list(c(x, params)),{
10     N <- S+I+R
11     dS <- -(beta*S*I)/N
12     dI <- (beta*S*I)/N - gamma*I
13     dR <- gamma*I
14     dx <- c(dS, dI, dR)
15     return(list(dx))
16   })
17 }
18 #3. Initial condition
19 N <- 1000 # population
20 S0 <- N - 1
21 I0 <- 1
22 R0 <- 0

```

(continued)

```

23 #4. Do the simulation (solving ODE system)
24 params <- c(beta = 0.5, gamma = 0.2)
25 initial_state <- c(S = S0, I = I0, R = R0)
26 times <- 0:200
27 model <- ode(initial_state, times, SIR, params)
28 # 5. Visualization
29 summary(model) # get the statistics summary of the model
30 SIR_data <- as.data.frame(model) # make a dataframe of model
   output
31 plot3 <- ggplot(data= SIR_data) +
32   geom_line(data = SIR_data, aes(x = time, y = I), color = "black",
33             lwd = 1)+ 
34   geom_line(data = SIR_data, aes(x = time, y = S), color = "red",
35             lwd = 1)+ 
36   geom_line(data = SIR_data, aes(x = time, y = R), color = "blue",
37             lwd = 1)+ 
38   labs(color = "", title = "SIR model", y = "\n" , x = "Time(
39     days)") +
40   theme(plot.title = element_text(size=10))+ 
41   theme(legend.position = "right")
42 plot3

```

Listing 1.3 R example

1.7.4 MATLAB Code

```

1 %% Set the parameter values
2 beta = 0.5; % mean transmission rate
3 gamma = 0.2; % mean recovery rate
4 %% Next Set the initial conditions
5 N = 1000; % total population size
6 I0 = 1; % initial # of infected individuals
7 S0 = N-1; % initial # of susceptible individuals
8 R0 = 0; % initial # of recovered individuals
9 initials = [S0 I0 R0]; % initial condition for integration of ode
10 %% Next solve the ode
11 period = 0:0.1:200; % period of interested
12 options = odeset('RelTol',1e-8,'AbsTol',1e-10); % set error
   tolerance
13 [time, y] = ode45(@(t,x) SIR(t,x,beta,gamma), period, initials,
   options);
14 %% Next Plot the epidemics
15 S = plot(time, y(:,1),'-', 'color', [1 0 0], 'linewidth',3);
16 hold on
17 I = plot(time, y(:,2),'-', 'color', [0 0 1], 'linewidth',3);
18 R = plot(time, y(:,3),'-', 'color', [0 1 0], 'linewidth',3);
19 hold off

```

(continued)

```

20 %% Next set the axis and graph labels
21 xlabel('time (days)', 'interpreter', 'latex', 'FontSize', 22);
22 ylabel('individuals', 'interpreter', 'latex', 'FontSize', 22);
23 title('SIR Model', 'interpreter', 'latex')
24 legend({'susceptible', 'infected', 'recovered'}, 'interpreter', 'latex')
25 set(gca, 'LineWidth', 1.3, 'FontSize', 22)
26 set(gcf, 'Color', 'white')
27 box off
28 %% Define your model
29 function dx = SIR(t, x, beta, gamma)
30     dx = zeros(3, 1);
31     N = x(1) + x(2) + x(3); % total population size
32     dx(1) = -beta * x(1) * x(2) / N; % susceptible population dynamics
33     dx(2) = beta * x(1) * x(2) / N - gamma * x(2); % infected population dynamics
34     dx(3) = gamma * x(2); % recovered population dynamics
35 end

```

Listing 1.4 Matlab example

1.8 Human Epidemiology Data, Model Fitting, and Parameter Estimation

Human data are the most desirable for use in the parameterization of models and are highly prioritized as they avoid the concern for species differences in response to exposures. Unfortunately, reliable epidemiological data are often unavailable or incomplete or contain unreliable exposure histories. For this reason, it is challenging to construct a valid and reliable mathematical model based on epidemiology studies. More often, the human studies can only provide qualitative evidence that a causal relationship exists. The basis for *sufficient human evidence* is an epidemiology study that clearly demonstrates a causal relationship between exposure and disease in humans. Data are determined to be *limited evidence in humans* if there are alternative explanations for the observed effect. The data are considered to be *inadequate evidence in humans* if no satisfactory epidemiology studies exist. For better predictions, more data is needed to refine the models being used. For example, it may be possible to decide optimal allocation of resources for treatment from a model when there are enough data to know susceptibility to infection for several different age groups [3].

Building a model that describes the transmission dynamics of an infectious disease will strongly depend on parameters and enough data to make proper estimation of unknown parameters and predictions. Nevertheless, this procedure comes with some fundamental challenges since models are based on unobservable occurrences at the time of modeling, such as the transmission of infection between infectives and

susceptible individuals, the start and end of an infectious period (the unexplained scenario, U in Fig. 1.1a [26]), and the serial interval (the time interval between sequential infectious individual in a series of transmission). However, data that are based on observable occurrences are usually collected by means of epidemiological and clinical evidence [3]. The clinical serial interval may differ from the serial interval from a model. Also challenging are differences in terminology, as public health professionals use the word *incubation period* of an infection (the time from the period of infection/exposure to the clinical onset of the disease as in Fig. 1.1a [3, 26]), while modelers use the word *latent/exposed period* (the time from the period of infection/exposure to the period of being infectious as in Fig. 1.1a [3, 26]). Inconsistent or inappropriate use of these concepts may lead to confusion and may not be appropriately accounted for in the model. An example is the case of influenza where there exists an infectious pre-symptomatic period, resulting in a shorter latent period than the incubation period. In this situation, individuals become infectious before showing symptoms and is denoted as U (unexplained) in Fig. 1.1a.

Another problem is the bias that may arise from data collection [3]. Administrative factors such as delays in reporting (report bias) and misclassification bias (inconsistencies in classifying clinical cases) may distort and complicate the analysis of clinical data. A disease such as influenza which has an infectious pre-symptomatic period and is therefore undiagnosed or not reported, and/or differences in reporting from one location to another, may experience a complicated or distorted clinical data analysis [3]. Data collected from an epidemic are commonly used to estimate key transmission parameters such as the basic reproduction number. This may be achieved using the observed initial exponential growth rate of infectious cases. Measuring the initial exponential growth rate (Υ) in Eq. (1.8) makes it easier to estimate the basic reproduction number (R_0). Nevertheless, it is unreasonable to fit curves to data if the model does not produce a curve that has the same distribution as the observed data [3]. In addition, there is also a problem of differences in *reported cases* (symptomatic cases) and *actual cases* (include both symptomatic and asymptomatic cases) of infection. The curve produced from epidemic data represents the reported cases, while a model often produces a curve that represents the actual cases of infection. Proper distinctions between these is necessary to obtain meaningful results.

The addition of an *exposed/latent period* (as in the *SEIR* model) will change the relationship between the initial exponential growth rate and the basic reproduction number [3, 13]. Therefore, the use of simplified model can lead to incorrect estimates of important parameters [3]. Some limitations such as a balance between predictive power of the model, its level of complication, and the type of questions to be addressed are inherent to the model structure itself. We therefore need to determine which parameters need to be included or excluded from the model based on their relevance and effect on the accuracy of predictions [70]. The accuracy of the data used for estimating parameters of the model determines how useful a model will be [3, 40]. In the case of limited data, *sensitivity and uncertainty analyses* may be conducted to determine the most important information for reliable estimate of outcomes [3, 40, 68]. Uncertainty analysis is done to investigate the effect of unknown parameters or missing data on model outputs, while sensitivity analysis is

done to investigate how model outputs vary with changes in input parameter values [3, 68]. These two methods are now commonly used in decision analysis and are now being used in infectious disease modeling. These methods help to identify parameter values that most influence model estimates and hence are needed with urgency [3, 68].

1.9 Conclusion

This chapter describes the impact of mathematical models on policy-making and the effect of available data on parameter estimation. We can draw several conclusions from this chapter and the study of infectious diseases in general. While data collected early in an epidemic often lack the validity and reliability needed to develop useful disease models, timely and accurate data are needed to develop models that compare management policies for disease outbreaks. In the face of uncertainty with respect to model parameters, uncertainty and sensitivity analyses can be used to determine which parameters are most likely to impact model projections.

To inform public health planning, disease control policy, and public health decision-making, quantitative modeling strategies require input from the collaboration of experts from different disciplines such as clinicians, public health professionals, laboratory technologists, epidemiologists, and mathematical modelers. Knowledge translation is a crucial part of modeling, and therefore, modelers need to include knowledge translation activities to demonstrate and communicate the relevance of their results in plain language and within the context of public health.

Acknowledgments At the time of revising this chapter, J.D. was a Public Health Analytics and Modeling Fellow at the US Centers for Disease Control and Prevention (CDC) and enjoyed the support of the CDC Prevention Effectiveness Fellowship; P.S. was supported by the Postdoctoral Fellowship of York University, Toronto, Canada, China Postdoctoral Science Foundation (No. 2020M683445), and the National Natural Science Foundation of China (NSFC, 12101487 (PS)); J.W. was supported by the Canada Research Chair Program (No. 105588-2011, 230720 (JW)). Authors acknowledged the comments by reviewers that improved the chapter.

References

1. Anderson, R.M.: The role of mathematical models in the study of hiv transmission and the epidemiology of aids. *J. Acquir. Immune Defic. Syndr.* **1**(3), 241–256 (1988)
2. Andrieu, C., Doucet, A., Holenstein, R.: Particle markov chain monte carlo methods. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **72**(3), 269–342 (2010)
3. Arino, J., Bauch, C., Brauer, F., Driedger, S.M., Greer, A.L., Moghadas, S.M., Pizzi, N.J., Sander, B., Tuite, A., Van Den Driessche, P., et al.: Pandemic influenza: modelling and public health perspectives. *Math. Biosci. Eng.* **8**(1), 1 (2011)

4. Arino, J., Driessche, P.v.d.: The basic reproduction number in a multi-city compartmental epidemic model. In: Positive Systems, pp. 135–142. Springer, New York (2003)
5. Ashish R. Hota Tanya Sneh, K.G.: Impacts of game-theoretic activation on epidemic spread over dynamical networks. *SIAM J. Control Optim.* **60**(2), S92–S118 (2022)
6. Aya, A., Aldila, D., Handari, B.: SDE model of SARS disease in Hong Kong and Singapore with parameter stochasticity. In: T. Mart (ed.) Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Sciences 2017 (ISCPMS 2017), p. 020218. AIP Conference Proceedings (2017)
7. Bélanger, A., Sabourin, P.: Microsimulation and Population Dynamics: An Introduction to Modgen 12. Springer, Cham, Switzerland (2017)
8. Bichara, D., Kang, Y., Castillo-Chavez, C., Horan, R., Perrings, C.: Sis and sir epidemic models under virtual dispersal. *Bull. Math. Biol.* **77**(11), 2004–2034 (2015)
9. Booth, T.F., Kournikakis, B., Bastien, N., Ho, J., Kobasa, D., Stadnyk, L., Li, Y., Spence, M., Paton, S., Henry, B., Mederski, B., White, D., Low, D.E., McGeer, A., Simor, A., Vearncombe, M., Downey, J., Jamieson, F.B., Tang, P., Plummer, F.: Detection of airborne severe acute respiratory syndrome (SARS) coronavirus and environmental contamination in SARS outbreak units. *J. Infect. Dis.* **191**, 1472–1477 (2005)
10. Brauer, F.: Age-of-infection and the final size relation. *Math. Biosci. Eng.* **5**(4), 681 (2008)
11. Brauer, F.: Mathematical epidemiology: Past, present, and future. *Infect. Dis. Model.* **2**(2), 113–127 (2017)
12. Brauer, F.: A new epidemic model with indirect transmission. *J. Biol. Dyn.* **11**(sup2), 285–293 (2017)
13. Brauer, F., Castillo-Chavez, C., Castillo-Chavez, C.: Mathematical Models in Population Biology and Epidemiology, vol. 2. Springer, New York (2012)
14. Campbell, J.: SARS Commission Final Report: Spring of Fear. Commission to Investigate the Introduction and Spread of SARS in Ontario (2006). www.sarscommission.ca
15. Casagrandi, R., Bolzoni, L., Levin, S.A., Andreasen, V.: The SIRC model and influenza a. *Math. Biosci.* **200**(2), 152–169 (2006)
16. Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore y Piontti, A., Mu, K., Rossi, L., Sun, K., et al.: The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science* **368**(6489), 395–400 (2020)
17. Choi, B., Pak, A.: A simple approximate mathematical model to predict the number of severe acute respiratory syndrome cases and deaths. *J. Epidemiol. Community Health* **57**, 831–835 (2003)
18. Chowell, G., Brauer, F.: The basic reproduction number of infectious diseases: computation and estimation using compartmental epidemic models. In: Mathematical and statistical estimation approaches in epidemiology, pp. 1–30. Springer, New York (2009)
19. Chowell, G., Fenimore, P., Castillo-Garsow, M., C., C.C.: SARS outbreak in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism. *J. Theor. Biol.* **224**, 1–8 (2003)
20. David, J., Bragazzi, N.L., Scarabel, F., McCarthy, Z., Wu, J.: Non-pharmaceutical intervention levels to reduce the covid-19 attack ratio among children. *R. Soc. Open Sci.* **9**(3), 211863 (2022)
21. Eaton, J.W., Bacaër, N., Bershteyn, A., Cambiano, V., Cori, A., Dorrington, R.E., Fraser, C., Gopalappa, C., Hontelez, J.A.C., Johnson, L.F., Klein, D.J., Phillips, A.N., Pretorius, C., Stover, J., Rehle, T.M., Hallet, T.B.: Assessment of epidemic projections using recent HIV survey data in South Africa: a validation analysis of ten mathematical models of HIV epidemiology in the antiretroviral therapy era. *Lancet Glob. Health* **3**, e598–608 (2015)
22. Eaton, J.W., Johnson, L.F., Salomon, J.A., Bärnighausen, T., Bendavid, E., Bershteyn, A., Bloom, D.E., Cambiano, V., Fraser, C., Hontelez, J.A., et al.: Hiv treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on hiv incidence in South Africa. *PLoS Med.* **9**(7), e1001245 (2012)
23. Endo, A., Van Leeuwen, E., Baguelin, M.: Introduction to particle markov-chain monte carlo for disease dynamics modellers. *Epidemics* **29**, 100363 (2019)

24. Frazier, P.I., Cashore, J.M., Duan, N., Henderson, S.G., Janmohamed, A., Liu, B., Shmoys, D.B., Wan, J., Zhang, Y.: Modeling for covid-19 college reopening decisions: Cornell, a case study. *Proc. Natl. Acad. Sci.* **119**(2), e2112532119 (2022)
25. Gandon, S., Lion, S.: Targeted vaccination and the speed of sars-cov-2 adaptation. *Proc. Natl. Acad. Sci.* **119**(3), e2110666119 (2022)
26. Gordis, L.: Epidemiology: with student consult online access (2014)
27. Granich, R.M., Gilks, C.F., Dye, C., De Cock, K.M., Williams, B.G.: Universal voluntary hiv testing with immediate antiretroviral therapy as a strategy for elimination of hiv transmission: a mathematical model. *Lancet* **373**(9657), 48–57 (2009)
28. Greenwood, P.E., Gordillo, L.F.: Stochastic epidemic modeling. In: Mathematical and statistical estimation approaches in epidemiology, pp. 31–52. Springer, New York (2009)
29. Hamer, W.H.: The Milroy Lectures on Epidemic Diseases in England: The Evidence of Variability and of Persistency of Type; Delivered Before the Royal College of Physicians of London, March 1st, 6th, and 8th, 1906. Bedford Press, Lambertville (1906)
30. Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars-Larrive, A., Loreto, V., Pinior, B., Thurner, S., Klimek, P.: Ranking the effectiveness of worldwide covid-19 government interventions. *Nat. Hum. Behav.* **4**(12), 1303–1312 (2020)
31. Hellewell, J., Abbott, S., Gimma, A., Bosse, N.I., Jarvis, C.I., Russell, T.W., Munday, J.D., Kucharski, A.J., Edmunds, W.J., Sun, F., et al.: Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *Lancet Glob. Health* **8**(4), e488–e496 (2020)
32. Hunter, E., Kelleher, J.: Framework for validating and testing agent-based models: a case study from infectious disease modelling. In: 34th Annual European Simulation and Modelling Conference. Toulouse, France (2020). <https://doi.org/doi:10.21427/2xjb-cq79>
33. Hunter, E., Mac Namee, B., Kelleher, J.: An open-data-driven agent-based model to simulate infectious disease outbreaks. *PloS One* **13**(12), 1–35 (2018). <https://doi.org/10.1371/journal.pone.0208775>
34. Hunter, E., MacNamee, B., Kelleher, J.: A taxonomy for agent-based models in human infectious disease epidemiology. *J. Artif. Soc. Soc. Simul.* **20**(3), 2 (2017)
35. Hunter, E., MacNamee, B., Kelleher, J.: Using a socioeconomic segregation burn-in model to initialise an agent-based model for infectious diseases. *J. Artif. Soc. Soc. Simul.* **21**(4), 9 (2018)
36. Jaber-Douraki, M., Moghadas, S.M.: Optimal control of vaccination dynamics during an influenza epidemic. *Math. Biosci. Eng.* **11**(5), 1045 (2014)
37. Jiang, H., Song, P., Wang, S., Yin, S., Yin, J., Zhu, C., Cai, C., Xu, W., Li, W.: Quantitative assessment of the effectiveness of joint measures led by fangcang shelter hospitals in response to covid-19 epidemic in Wuhan, China. *BMC Infect. Dis.* **21**(1), 1–11 (2021)
38. JUNPo, H.I.V., AIDS HIV.: Aids JUNPo: 90–90–90: an ambitious treatment target to help end the AIDS epidemic. Geneva: UNAIDS (2014)
39. Karatayev, V.A., Anand, M., Bauch, C.T.: Local lockdowns outperform global lockdown on the far side of the covid-19 epidemic curve. *Proc. Natl. Acad. Sci.* **117**(39), 24575–24580 (2020)
40. Keeling, M., Danon, L.: Mathematical modelling of infectious diseases. *Br. Med. Bull.* **92**(1), 33–42 (2009)
41. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A Contain. Papers Math. Phys. Char.* **115**(772), 700–721 (1927)
42. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics. ii.–the problem of endemicity. *Proc. R. Soc. Lond. Ser. A Contain. Papers Math. Phys. Char.* **138**(834), 55–83 (1932)
43. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics. ii.–the problem of endemicity. *Proc. R. Soc. Lond. Ser. A Contain. Papers Math. Phys. Char.* **141**(834), 94–112 (1932)
44. Kim, Y., Ryu, H., Lee, S.: Agent-based modeling for super-spreading events: A case study of MERS-CoV transmission dynamics in the Republic of Korea. *Int. J. Environ. Res. Public Health* **15**(11), 2369 (2018)

45. Kucharski, A.J., Klepac, P., Conlan, A.J., Kissler, S.M., Tang, M.L., Fry, H., Gog, J.R., Edmunds, W.J., Emery, J.C., Medley, G., et al.: Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of sars-cov-2 in different settings: a mathematical modelling study. *Lancet Infect. Dis.* **20**(10), 1151–1160 (2020)
46. Lei, H., Li, Y., Xiao, S., Lin, C., Norris, S.L., Wei, D., Hu, Z., Ji, S.: Routes of transmission of influenza A H1N1, SARS CoV, and norovirus in air cabin: Comparative analyses. *Indoor Air* **28**, 394–403 (2018)
47. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y., et al.: Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020)
48. Li, u., Yu, I.T.S., Xu, P., Lee, J.H.W., Wong, T.W., Ooi, P.L., Sleigh, A.C.: Predicting super spreading events during the 2003 severe acute respiratory syndrome epidemics in Hong Kong and Singapore. *Am. J. Epidemiol.* **160**, 719–728 (2004)
49. Li, Y., Huang, X., Yu, I., Wong, T., Qian, H.: Role of air distribution in SARS transmission during the largest nosocomial outbreak in Hong Kong. *Indoor Air* **15**, 83–95 (2004)
50. Lima, V.D., Johnston, K., Hogg, R.S., Levy, A.R., Harrigan, P.R., Anema, A., Montaner, J.S.: Expanded access to highly active antiretroviral therapy: a potentially powerful strategy to curb the growth of the hiv epidemic. *J. Infect. Dis.* **198**(1), 59–67 (2008)
51. Lima, V.D., Rozada, I., Grebely, J., Hull, M., Lourenco, L., Nosyk, B., Krajden, M., Yoshida, E., Wood, E., Montaner, J.S.: Are interferon-free direct-acting antivirals for the treatment of hcv enough to control the epidemic among people who inject drugs? *PloS One* **10**(12), e0143836 (2015)
52. Lin Wang, G.Z.D.: Global stability of virus spreading in complex heterogeneous networks. *SIAM J. Appl. Math.* **68**(5), 1495–1502 (2008)
53. Lucas M. Stolerman Daniel Coombs, S.B.: Sir-network model and its application to dengue fever. *SIAM J. Appl. Math.* **75**(6), 2581–2609 (2015)
54. Martin, N.K., Vickerman, P., Grebely, J., Hellard, M., Hutchinson, S.J., Lima, V.D., Foster, G.R., Dillon, J.F., Goldberg, D.J., Dore, G.J., et al.: Hepatitis c virus treatment for prevention among people who inject drugs: modeling treatment scale-up in the age of direct-acting antivirals. *Hepatology* **58**(5), 1598–1609 (2013)
55. Meyers, L.A., Pourbohloul, B., Newman, M.E., Skowronski, D.M., Brunham, R.C.: Network theory and sars: predicting outbreak diversity. *J. Theor. Biol.* **232**(1), 71–81 (2005)
56. Mills, C.E., Robins, J.M., Lipsitch, M.: Transmissibility of 1918 pandemic influenza. *Nature* **432**(7019), 904–906 (2004)
57. Montaner, J.S., Hogg, R., Wood, E., Kerr, T., Tyndall, M., Levy, A.R., Harrigan, P.R.: The case for expanding access to highly active antiretroviral therapy to curb the growth of the hiv epidemic. *Lancet* **368**(9534), 531–536 (2006)
58. Montaner, J.S., Lima, V.D., Harrigan, P.R., Lourenço, L., Yip, B., Nosyk, B., Wood, E., Kerr, T., Shannon, K., Moore, D., et al.: Expansion of haart coverage is associated with sustained decreases in hiv/aids morbidity, mortality and hiv transmission: the “hiv treatment as prevention” experience in a canadian setting. *PloS One* **9**(2), e87872 (2014)
59. Nah, K., Nishiura, H., Tsuchiya, N., Sun, X., Asai, Y., Imamura, A.: Test-and-treat approach to hiv/aids: a primer for mathematical modeling. *Theor. Biol. Med. Model.* **14**(1), 1–11 (2017)
60. Naheed, A., Singh, M., Lucy, D.: Numerical study of SARS epidemic model with the inclusion of diffusion in the system. *Appl. Math. Comput.* **229**, 480–498 (2014)
61. Ojo, M.M., Benson, T.O., Peter, O.J., Goufo, E.F.D.: Nonlinear optimal control strategies for a mathematical model of covid-19 and influenza co-infection. *Physica A Stat. Mech. Appl.* **607**, 128173 (2022)
62. Potter, C.W.: A history of influenza. *J. Appl. Microbiol.* **91**(4), 572–579 (2001)
63. Pourbohloul, B., Miller, J.: Network theory and the spread of communicable diseases. *Center Dis. Model. Preprint* **3**, 4–16 (2008)
64. Rafferty, E., McDonald, W., Qian, W., Osgood, N., Doroshenko, A.: Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling. *PeerJ* **6**, e5012 (2018)

65. Romualdo Pastor-Satorras, A.V.: Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001)
66. Ross, R.: *The Prevention of Malaria*. John Murray, London (1911)
67. Roy, C.J., Milton, D.K.: Airborne transmission of communicable infection—the elusive pathway. *N. Engl. J. Med.* **350**(3), 1710–1712 (2004)
68. Rozada, I., Coombs, D., Lima, V.D.: Conditions for eradicating hepatitis c in people who inject drugs: A fibrosis aware model of hepatitis c virus transmission. *J. Theor. Biol.* **395**, 31–39 (2016)
69. Rvachev, L.A., Longini Jr, I.M.: A mathematical model for the global spread of influenza. *Math. Biosci.* **75**(1), 3–22 (1985)
70. Star, L., Moghadas, S.: The role of mathematical modelling in public health planning and decision making. Purple Paper, National Collaborative Center for Infectious Diseases (2010)
71. Tang, B., Wang, X., Li, Q., Bragazzi, N.L., Tang, S., Xiao, Y., Wu, J.: Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions. *J. Clin. Med.* **9**(2), 462 (2020)
72. Taubenberger, J.K., Morens, D.M.: 1918 influenza: the mother of all pandemics. *Revista Biomedica* **17**(1), 69–79 (2006)
73. Valleron, A.J., Cori, A., Valtat, S., Meurisse, S., Carrat, F., Boëlle, P.Y.: Transmissibility and geographic spread of the 1889 influenza pandemic. *Proc. Natl. Acad. Sci.* **107**(19), 8778–8781 (2010)
74. Van den Driessche, P., Watmough, J.: Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **180**(1–2), 29–48 (2002)
75. Wang, L., Liu, Z., Xu, D., Zhang, X.: Global dynamics and optimal control of an influenza model with vaccination, media coverage and treatment. *Int. J. Biomath.* **10**(05), 1750068 (2017)
76. Wang, W., Ruan, S.: Simulating the sars outbreak in beijing with limited data. *J. Theor. Biol.* **227**(3), 369–379 (2004)
77. Webb, G., Blaser, M.J., Zhu, H., Ardal, S., Wu, J.: Critical role of nosocomial transmission in the toronto sars outbreak. *Math. Biosci. Eng.* **1**(1), 1 (2004)
78. Weitz, J.S., Park, S.W., Eksin, C., Dushoff, J.: Awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations. *Proc. Natl. Acad. Sci.* **117**(51), 32764–32771 (2020)
79. Wilasang, C., Suttirat, P., Chadsuthi, S., Wiratsudakul, A., Modchang, C.: Competitive evolution of h1n1 and h3n2 influenza viruses in the united states: A mathematical modeling study. *J. Theor. Biol.*, 111292 (2022)
80. Wilensky, U.: NetLogo (and NetLogo User Manual). Center for Connected Learning and Computer-Based Modeling. Northwestern University (1999). <http://ccl.northwestern.edu/netlogo/>
81. Wolfson, M.: POHEM: a framework for understanding and modelling the health of human populations. *World Health Stat. Q.* **47**, 157–176 (1994)
82. World Health Organization, et al.: Pandemic Influenza Preparedness and Response: A WHO Guidance Document. World Health Organization (2009)
83. Yip, C., Chang Wen, L., Yeung, K.H., Yu, I.T.S.: Possible meteorological influence on the severe acute respiratory syndrome (SARS) community outbreak at Amoy Gardens, Hong Kong. *J. Environ. Health* **70**(3), 39–47 (2007)
84. Yu, I., Li, Y., Wong, T., Tam, W., Chan A.T.and Lee, J., Leung, D., Ho, T.: Evidence of airborne transmission of the severe acute respiratory syndrome virus. *N. Engl. J. Med.* **350**, 1731–1739 (2004)
85. Zhou, Y., Ma, Z., Brauer, F.: A discrete epidemic model for SARS transmission and control in China. *Math. Comput. Model.* **40**, 1491–1506 (2004)

Chapter 2

Discovering First Principle of Behavioural Change in Disease Transmission Dynamics by Deep Learning



Pengfei Song, Yanni Xiao, and Jianhong Wu

2.1 Introduction

Infectious diseases can have a large impact on society as they can negatively affect, among others, morbidity, mortality, unemployment and inequality. The ongoing COVID-19 pandemic since first officially reported in Wuhan, China, in late 2019, poses continuing threat on human's health [30]. Controlling emerging viral infectious disease will depend critically on non-pharmaceutical interventions including social distancing, mask wearing, contact tracing, isolation, quarantine and border control and pharmaceutical measures such as vaccination and antiviral drugs [3]. The successful implementation of public health interventions is greatly subject to human behaviours as the key determinant of the course, duration and outcomes of disease outbreaks [2, 47, 48, 53]. Therefore, there is a great interest to incorporate behavioural change in response to disease-related information into models for infectious disease transmission dynamics.

Behavioural change transmission dynamics models share a tremendous popularity in recent years; see reviews [18, 44, 47] for more details. To characterize the human behavioural change in transmission dynamics models, two ways are common. The first way is to add more compartments representing risk-aware and risk-unaware subpopulations; see [16, 17] for example. The second way is to

P. Song · J. Wu (✉)

Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics,
York University, Toronto, ON, Canada

e-mail: song1012@york.ca; wujh@york.ca

Y. Xiao

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, PRC
e-mail: yxiao@mail.xjtu.edu.cn

change the incidence rate by including a function which decreases as infection increases. For example, [31] introduced a media function $\beta e^{-\alpha_1 E - \alpha_2 I - \alpha_3 H}$ into the transmission coefficient, where E , I and H are the numbers of exposed, infectious and hospitalized individuals, respectively. The study [49] extended these media functions by assuming that the function depends on both the case number and its rate of change and obtained that media impact switches on and off in a highly nonlinear fashion. The subsequent work [50] further extended these models by including extra compartment, i.e. the level of media coverage M , and consequently media impact is modelled by including the function $e^{-\mu M}$ with $\mu > 0$ in the incidence.

However, these behavioural change transmission dynamics models are heavily experts-based or assumptions-based. The successful applications of these models in epidemic control are greatly dependent on good assumptions. Simple assumptions make transmission dynamics models theoretically tractable, but are not prone to fit the real epidemic data; complicated assumptions can fit the real epidemic data better, but often make transmission dynamics computations intensive and may not have good generalization and predictability. How to characterize the behavioural changes in transmission dynamics model with a simpler and more data-driven way is still a challenge. Behavioural change models must strike a delicate balance, harnessing available data and theory on complex social and behavioural phenomena while keeping important modelling properties, such as computational tractability and simplicity.

In recent years, deep neural networks [19, 29], as a universal approximator for unknown mappings [23, 27] and nonlinear operators [32], show incredible effectiveness [41] in pattern recognition, learning unknown mechanisms and solving traditional difficult problems such as image recognition [20] and natural language processing [14]. Recently, deep learning methods have been used to aid mathematicians in discovering new conjectures and theorems [12] and physicians in finding new physical laws [11]. Moreover, to better understand and improve the performance of deep neural networks on scientific problems in physics, chemistry and epidemiology, increasing attentions have been paid to coupling or embedding differential equations and deep neural network. One important idea is regarding deep neural network (DNN) as discretization of differential equation, which inspires researchers to redesign traditional sequential neural architectures based on numerical discretization schemes [9, 33, 34, 38] or to replace DNN by a continuous model characterized by differential equations [8, 13, 37]. Another revolutionary idea is training DNN from the view of optimal control. This extends the idea of backpropagation [19] to include adjoint sensitivity analysis [7], which opens the new research frontiers in differential programming [1] and inspires a lot of architectures coupling neural networks and differential equations, such as neural differential equation [8] and universal differential equation [35].

Although neural differential equations and universal differential equations allow efficient training of complex models on high-dimensional datasets and show incredible learning effectiveness, black-box terms are still unwelcome. One still

cares about what the exact equation should be and wishes the equation to be as simple as possible, rather than represented as a neural network and difficult to interpret in differential equations. Data-driven methods, like symbolic regression, sparse identification of nonlinear dynamical systems [5, 6, 25, 39] and dynamic mode decomposition [5, 15, 45], have been proposed in recent years for this purpose. Unlike deep neural networks, data-driven methods are interpretable equation-search methods with the purpose to find the simplest analytic formulas to describe science, engineering and real-world data. However, data-driven methods show the weakness in training efficiency in handling high-dimension differential equations, when compared to deep learning methods, and the dynamics are difficult to learn when data observation is limited to small number of samples.

In this work, we propose a two-step recovering-explaining framework which combines deep learning methods and data-driven methods and use the framework to discover the exact expression of the unknown behavioural change mechanisms in transmission dynamics models. We mention that the proposed framework balances the learning ability and interprecity of the transmission dynamics models with neural networks embedded and can be used to address other relevant issues where hidden mechanisms may be explored with limited epidemic data.

The rest of the paper is organized as follows: An expert-based human behavioural change disease transmission dynamics model is proposed in Sect. 2.2 to show that a good assumption for behavioural change transmission dynamics model is theoretically tractable and can have good predictability. We will introduce two-step recovering-explaining framework in Sect. 2.3. In Sect. 2.4, the framework is applied to Ontario COVID-19 data to discover the behavioural change mechanisms in the province. Discussions and conclusions will be in Sect. 2.5.

2.2 Expert-Based Behavioural Change Transmission Dynamics Models

In this section, we propose a novel simple SIR (suspected-infected-recovered) model incorporated with behavioural change mechanisms and show that the simple behavioural change SIR model is theoretically tractable and has good generalization and predictability in the first wave of COVID-19 pandemic in the province of Ontario, Canada. This simple behavioural change SIR model reveals the importance of good assumptions in expert-based models and necessities to use deep learning methods to discover unknown mechanisms and find good assumptions in transmission dynamics models.

We extend the classic SIR model [26] to include a compartment to describe the dynamics of human behaviours as follows:

$$\begin{cases} \frac{dS}{dt} = -\beta c SI, \\ \frac{dI}{dt} = \beta c SI - \gamma I, \\ \frac{dR}{dt} = \gamma I, \\ \frac{d \ln c}{dt} = -\theta \frac{d \ln I}{dt} - \delta I, \end{cases} \quad (2.1)$$

where S and I denote the number of susceptible and infectious individuals, respectively, β is the baseline infection rate, γ is the recovery rate, and c is the average number of disease transmission effective contacts to describe the human behavioural change effect. Here we assume that the changing rate of human behaviours depends on the prevalence and changing rate of prevalence.

2.2.1 Calculation of the Final Epidemic Size

To start with, by using

$$I = \frac{1}{\gamma} \frac{dR}{dt},$$

the third equation can be written as

$$\frac{d(\ln c + \theta \ln I + \delta R/\gamma)}{dt} = 0,$$

which implies

$$\ln(cI^\theta) + \delta R/\gamma = \ln(c_0 I_0^\theta) + \delta R_0/\gamma := M_0,$$

i.e.

$$c = I^{-\theta} \exp(M_0 - \delta R/\gamma).$$

Thus, the equation becomes

$$\begin{cases} \frac{dS}{dt} = -\beta \exp(M_0 - \delta R(t)/\gamma) SI^{1-\theta}, \\ \frac{dI}{dt} = \beta \exp(M_0 - \delta R(t)/\gamma) SI^{1-\theta} - \gamma I, \\ \frac{dR}{dt} = \gamma I. \end{cases}$$

By using $I = \frac{1}{\gamma} \frac{dR}{dt}$, the first equation becomes

$$\frac{d \ln S}{dt} = -\beta \exp(M_0 - \delta R(t)/\gamma) \left(\frac{1}{\gamma} \frac{dR}{dt} \right)^{1-\theta}.$$

Denote the final epidemic size as r . It can be easily proved that as $t \rightarrow \infty$,

$$S \rightarrow 1 - r, \quad I \rightarrow 0, \quad R \rightarrow r.$$

If $\theta = 0$, then

$$\frac{d \ln S}{dt} = \frac{\beta}{\delta} \left(\frac{d(\exp(M_0 - \delta R(t)/\gamma))}{dt} \right),$$

which implies

$$\ln(1 - r) = \frac{\beta}{\delta} \exp(M_0 - \delta r/\gamma) - \ln c_0.$$

If $\delta = 0$, we have

$$\frac{d \ln S}{dt} = -\beta c_0 I_0^\theta \left(\frac{1}{\gamma} \frac{dR}{dt} \right)^{1-\theta}.$$

In summary, we have the following theorem:

Theorem 1 Denote the final epidemic size of the behavioural change disease transmission dynamics model (2.1) as r . We have the following:

- (i) If $\theta = 0$ and $\delta \neq 0$, i.e. changing rate of human behaviours only depends on prevalence, then the final epidemic size satisfies

$$\ln(1 - r) = \frac{\beta}{\delta} \exp(M_0 - \delta r/\gamma) - \ln c_0.$$

- (ii) If $\theta \neq 0$ and $\delta = 0$, i.e. the change rate of human behaviours only depends on the changing rate of prevalence, then the final epidemic size satisfies

$$\frac{d \ln S}{dt} = -\beta c_0 I_0^\theta \left(\frac{1}{\gamma} \frac{dR}{dt} \right)^{1-\theta}.$$

Remark We note from (i) that if $\theta = \delta = 0$, i.e. without behavioural changes, then the final epidemic size satisfies

$$\ln(1 - r) = -R_0 r, \quad R_0 = \frac{\beta c_0}{\gamma}.$$

2.2.2 Applications to the Ontario's First COVID-19 Pandemic Wave

In this part, model (2.1) is applied on the Ontario's first COVID-19 pandemic wave. In the province, the outbreak began on February 25, 2020, and the first wave continued for about 150 days. The case data is shown in Fig. 2.1.

To start with, we use a simple SIR model to investigate the first wave data; we find that even with full 150 days' training data, a simple SIR model cannot fit well, which implies a simple SIR model without considering human behavioural changes is not enough to capture the evolution of epidemic. The results are shown in Fig. 2.2.

However, by using the novel behavioural change model, we find that using only 50 days of data is enough to predict trend of the epidemic in the first wave; see Fig. 2.1 for details. Here models are calibrated by non-U-turn Hamiltonian Monte Carlo method [21], which is implemented in Julia 1.7.3, an open-source software. We kept the same range of parameter values when we fit the models without and with behavioural change. More details on parameter values, implementation of parameter estimation and confidence intervals can be seen in [Github repo](#).

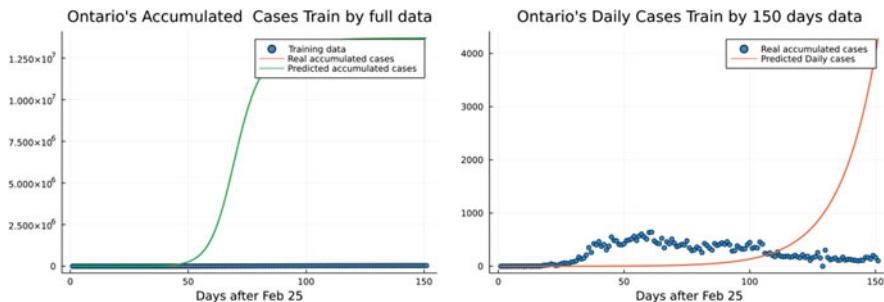


Fig. 2.1 Failures of simple SIR model to fit first wave Ontario COVID-19 even with full observation data

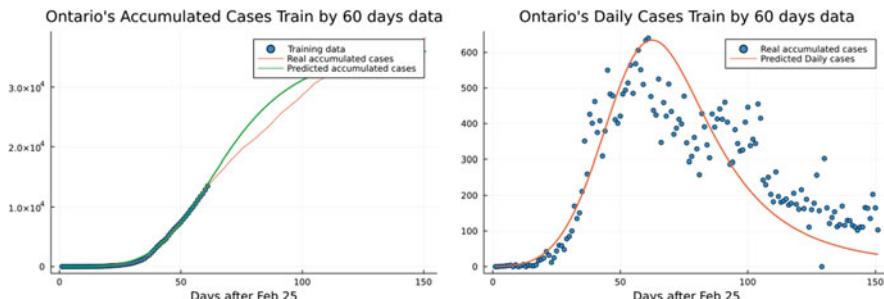


Fig. 2.2 Calibration of a transmission dynamics model incorporating behavioural changes to the 150 days of incidence data during the first wave of COVID-19 in the province of Ontario, where only the first 60 days of the data is used to produce a good prediction for the remaining 90 days of the first wave

2.3 Two-Step Recovering-Explaining Framework

For expert-based transmission dynamics models, describing unknown mechanisms, such as human behavioural changes, greatly depends on intuitions and good assumptions. However, finding a good assumption to describe unknown mechanisms in dynamic systems is a challenge. Here we propose a deep learning method-based two-step recovering-explaining framework to discover unknown mechanisms and make good assumptions.

2.3.1 Universal Differential Equations

We start with a short introduction of the state-of-the-art methodology, using universal differential equations (UDEs) [35], which embeds universal approximators in differential equations for scientifically based learning and can be used to discover previously unknown mechanisms and accurately extrapolate beyond the original data. Before the introduction on UDEs, we introduce neural differential equations [8] first, which were proposed before UDEs, inspired the ideas in UDEs and can be regarded as a special case of UDEs. Neural differential equations are initial value problems with the following form:

$$u' = \text{NN}_\theta(u, t), \quad (2.2)$$

where the right-hand side NN is a deep neural network receiving $[u, t]$ as input. Neural differential equations make use of scientific structures as a modelling basis and because the embedded function is a universal approximator. Thus, neural differential equations can learn to approximate any sufficiently regular differential equation. From the perspective of deep learning, neural differential equations are redesigned sequential neural networks based on numerical schemes of differential equation, and they can be regarded as continuous-depth or “infinitely deep” ResNet-like deep learning models. From the perspective of optimal control, the training of neural differential equations can be regarded as optimal control problems, which extend the idea of backpropagation [19] in differential programming to include adjoint sensitivity analysis [7].

However, the resulting neural differential model is defined without direct incorporation of known mechanisms. UDEs extend the previous data-driven neural ODE approaches to directly utilize mechanistic modelling simultaneously with universal approximators in order to allow for arbitrary data-driven model extensions. UDEs are initial value problems with the following form:

$$u' = f_{\theta_2}(u, t, U A_{\theta_1}(u, t)), \quad (2.3)$$

where f is a known mechanistic model and UA denotes the missing or unknown terms which can be represented by universal approximators such as neural networks and Gaussian process. Throughout the rest of this study, we choose neural networks as universal approximator. θ_1 and θ_2 are parameters of known mechanisms and neural networks, respectively, which can be estimated simultaneously. UDEs have strong explainability than deep neural networks or neural differential equations, since they keep known mechanisms in physics, chemistry or epidemiology. UDEs have been proved to be methods with good generalization and can be trained with less sample data [35].

2.3.2 Data-Driven Methods or Equation-Searching Methods

Although neural differential equations and universal differential equations allow efficient training of complex models on high-dimensional datasets and show incredible learning effectiveness, black-box terms are still unwelcome. One still cares about what the exact equation should be and wishes the equation to be as simple as possible, rather than represented as neural network and difficult to interpret in differential equations. Data-driven methods like sparse identification of nonlinear dynamical systems [5, 6, 25, 39] and dynamic mode decomposition [5, 15, 45] have been proposed in recent years for this purpose. Unlike deep neural network, data-driven methods are interpretable equation-search methods with the purpose to find the simplest analytic expressions to describe science, engineering and real-world data. In what follows, we will give a brief introduction of two famous data-driven methods: symbolic regression and sparse identification of nonlinear dynamics (SINDy).

2.3.2.1 Symbolic Regression

Symbolic regression (SR) is a type of regression analysis that searches the space of mathematical expressions to find the model that best fits a given dataset, both in terms of accuracy and simplicity. SR uses binary tree to represent a function, and no particular model is provided as a starting point to the algorithm. Instead, initial expressions are formed by randomly combining mathematical building blocks such as:

- Binary mathematical operators: $+, -, *, /$
- Unary analytic functions: $\sin, \cos, \exp, \tanh, etc.$
- Constants
- State variables

Usually, a subset of these primitives will be specified by the person operating it, but that's not a requirement of the technique. SR uses genetic programming [40], as

well as more recently methods utilizing Bayesian methods [24] and deep learning methods [46] to discover the equations.

By not requiring a priori specification of a model, symbolic regression is not affected by human bias, or unknown gaps in domain knowledge. It attempts to uncover the intrinsic relationships of the dataset, by letting the patterns in the data itself reveal the appropriate models, rather than imposing a model structure that is deemed mathematically tractable from a human perspective. The fitness function that drives the evolution of the models takes into account not only error metrics (to ensure the models accurately predict the data) but also special complexity measures, thus ensuring that the resulting models reveal the data's underlying structure in a way that's understandable from a human perspective.

2.3.2.2 Sparse Identification of Nonlinear Dynamics (SINDy)

The SINDy (sparse identification of nonlinear dynamics) algorithm [6] provides a principled, data-driven discovery method for nonlinear dynamics of the form

$$u' = f(u) \quad (2.4)$$

where $u(t) = [u_1(t); u_2(t); \dots; u_n(t)] \in \mathbb{R}^n$ is system states represented as a row vector. SINDy applies a set of candidate functions that would characterize the right-hand side of the governing equations. Candidate model terms form the library

$$\Theta(\mathbf{U}) = [\theta_1(\mathbf{U}), \theta_2(\mathbf{U}), \dots, \theta_p(\mathbf{U})] \in \mathbb{R}^{m \times p}$$

of potential right-hand side terms, where

$$\mathbf{U} = \begin{bmatrix} u^T(t_1) \\ u^T(t_2) \\ \vdots \\ u^T(t_m) \end{bmatrix} = \begin{bmatrix} u_1(t_1) & u_2(t_1) & \cdots & u_n(t_1) \\ u_1(t_2) & u_2(t_2) & \cdots & u_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ u_1(t_m) & u_2(t_m) & \cdots & u_n(t_m) \end{bmatrix}$$

is an $\mathbb{R}^{m \times n}$ matrix. $\theta_i(\mathbf{U})$ can be any candidate basis function that may describe the system dynamics $f(u(t))$ such as trigonometric functions $\theta_i(\mathbf{U}) = \cos(\mathbf{U})$ or polynomial functions $\theta_i(\mathbf{U}) = \mathbf{U}^2$. This then allows for the formulation of a regression problem to select only the few candidate terms necessary to describe the dynamics:

$$\arg \min_{\Xi} \|\dot{\mathbf{U}} - \Theta(\mathbf{U})\Xi\|_2 + \lambda \|\Xi\|_0$$

where the matrix

$$\Xi = [\xi_1, \xi_2, \dots, \xi_n] \in \mathbb{R}^{p \times n}$$

is comprised of the sparse vectors $\xi_1 \in \mathbb{R}^p$ that select candidate model terms. The amount of sparsity promotion is controlled by the parameter λ , which determines the penalization by the ℓ_0 -norm. By solving system (2.4), we can identify a model of system dynamics

$$u' = f(u) \approx \Theta(u)\Xi. \quad (2.5)$$

2.3.3 Two-Step Recovering-Explaining Methods

However, when compared to deep learning methods, data-driven methods show the weakness in training efficiency in handling high-dimension differential equations, and the dynamics are difficult to learn when observed data is limited to small number of samples. Can we combine these two areas' methods? Two-step recovering-explaining framework can be an answer.

In epidemiology, researchers care about both learning ability and interpretability. We want to efficiently learn the missing mechanisms in transmission dynamics models, which implies UDEs [35] and physics-informed neural networks (PINNs; [36]) are good choices. However, we do not wish to represent the missing mechanism as a black-box neural network because it is difficult to interpret in epidemiology, which implies data-driven methods are good choices. Recall that observed data in epidemiology are noisy, partially observed, sparsely sampled, and with heterogeneity between datasets, which give a lot of challenges on data-driven methods such as SINDy [6]. Is there a way to balance the learning ability and interpretability, efficiently learn the missing mechanisms in transmission dynamics and represent the mechanisms as simple as possible? In what follows, we propose a two-step recovering-explaining method as a solution. The key points of two-step recovering-explaining methods are as follows:

- Firstly we use machine learning methods like UDEs [35] or PINNs [36] to recover the partial observed, sparsely sampled and noisy data to fully observed, continuous and differentiable data.
- Then we apply the recovered data to find the unknown simple equations in transmission dynamics by data-driven methods like symbolic regression [40] or SINDy [6].

For epidemic models, the observed data such as death, accumulated and daily confirmed cases are not exact solutions of the differential equations. The state in transmission dynamics models are often regarded as the latent state of the observed data. Thus, it is difficult to apply data-driven methods like SINDy to discover the unknown mechanisms. Deep learning methods based on differential programming do not need the case data in epidemiology to be fully observed and differentiable. Moreover, deep learning methods can be regarded as an advanced data smoothing method showing incomparable performance to other traditional approaches such as moving average. Data assimilation methods like particle filter

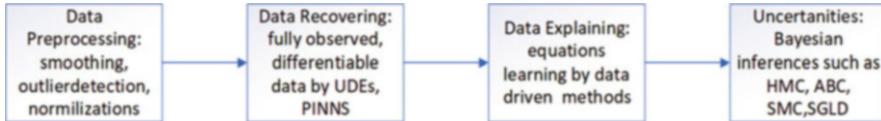


Fig. 2.3 A schematic illustration of the four major steps in the proposed two-step recovering-explaining method

[28] can recover fully observed, differentiable data, but knowledge on the unknown mechanisms is required a priori. Here we remark that two-step methods can also be trained simultaneously, if the algorithms in data-driven methods are differential programming based, for example, OccamNet [10].

The processes of two-step recovering-explaining methods are shown (Fig. 2.3) as follows:

- Step one: data reprocessing [4], such as data smoothing and outlier detection [51]
- Step two: data recovering—using machine learning methods like UDEs [35] or PINNs [36] to recover the partial observed, sparsely sampled and noisy data to fully observed, continuous and differentiable data
- Step three: data explaining—finding the unknown simple equations in transmission dynamics models by data-driven methods like symbolic regression [40] or SINDy [6]
- Step four: handling uncertainties—keeping the formula found by data-driven methods, using the parameters in the equations as prior knowledge and using Bayesian inference methods such as non-U-turn Hamiltonian Monte Carlo to handle the uncertainties

2.4 Deep Learning-Based Behavioural Change Transmission Dynamics Models

In this section, the two-step recovering-explaining framework is used to discover the unknown behavioural change mechanisms. In particular, we will use the case data from the first and second wave of the COVID-19 in Ontario to fit the following neural differential equation model:

$$\begin{aligned}
 \frac{dS}{dt} &= -\text{abs}(NN(I, R))S/N, \\
 \frac{dI}{dt} &= -\text{abs}(NN(I, R))S/N - \gamma I, \\
 \frac{dR}{dt} &= \gamma I,
 \end{aligned}$$

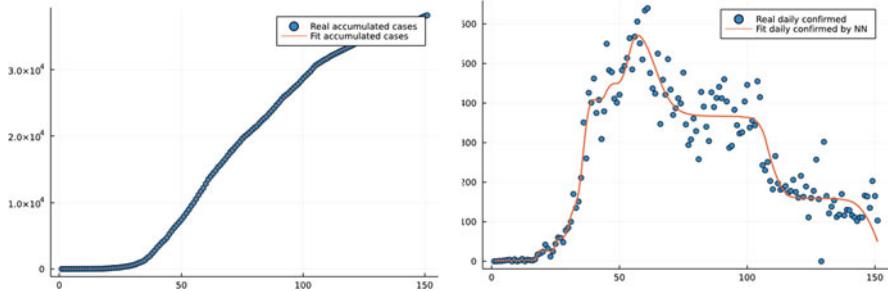


Fig. 2.4 Learn Ontario first wave data (150 days after February 25, 2020) by universal differential equations

$$\frac{dH}{dt} = \text{abs}(NN(I, R))S/N,$$

where $NN(I, R)$ denotes the neural network to learn the media impact function and H denotes the accumulated cases. Deep learning methods and symbolic regression methods are implemented in open-source Julia language 1.7.3 [54], Laptop Y7000P with i5-9300HF CPU, 16G RAM. The training time of deep neural networks and the training time of symbolic regression methods are approximately 1 hour and 30 seconds, respectively. All the details of the algorithms and codes can be found in [Github repo](#).

2.4.1 The Behavioural Change Laws

To start with, we recover the data by UDEs; the results are shown in Fig. 2.4.

After recovering the data, we use symbolic regression to find the simplest equation to fit $\text{abs}(NN(I, R))$, and the equation found is kind of saturated function

$$\text{abs}(NN(I, R)) \approx \frac{aI + b}{R + d},$$

which implies that

$$c' = -\frac{bc}{I(aI + b)}I' - \frac{\gamma I^2 c^2}{aI + b}.$$

The result is shown in Fig. 2.5.

Using the same framework, we also fit Ontario's second COVID-19 pandemic wave data and explore the evolution of human behaviour pattern. We first recover the data by UDEs; the results are shown in Fig. 2.6.

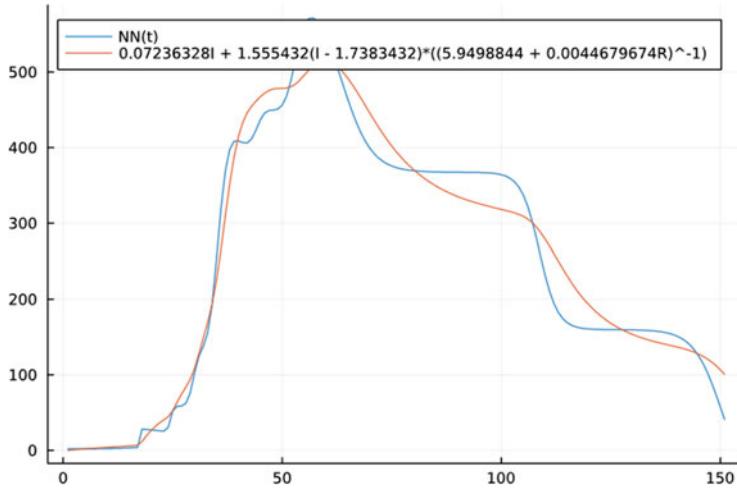


Fig. 2.5 Symbolic regression to find the simplest equation to fit $\text{abs}(NN(I, R))$

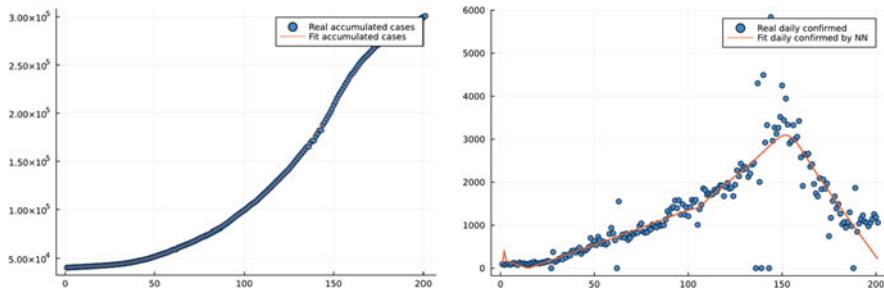


Fig. 2.6 Learn Ontario's second COVID-19 pandemic wave data (200 days after July 25, 2021) by universal differential equations

After recovering the data, we use symbolic regression to find the simplest equation to fit $\text{abs}(NN(I, R))$, and the equation found is as follows:

$$\text{abs}(NN(I, R)) \approx \left| \frac{aI}{(R + b)^c} - d \right|,$$

which implies that

$$c' = \frac{c}{I} + \frac{I^{1.004}(cI^{1.004} + 520)^2}{26} - \left(\frac{c}{I} + 520I^{0.004} \right)$$

The results are shown in Fig. 2.7.

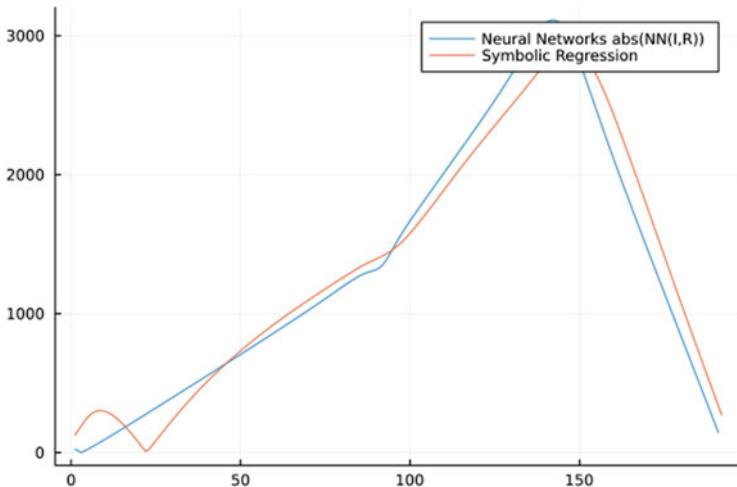


Fig. 2.7 Symbolic regression to find the simplest equation to fit $\text{abs}(NN(I, R))$ of Ontario’s second COVID-19 pandemic wave data

2.5 Discussions and Conclusions

Human behaviours can shape the spread of infectious disease and shift the epidemics away from peaks and toward plateaus, shoulders and oscillations. Study of human behavioural change disease transmission dynamics models can guide public health measures, such as reports of mass media. Nonetheless, the unknown behavioural change mechanisms restrict the ability of disease transmission dynamics in epidemic control. To characterize the human behaviour effect on infectious diseases, traditional epidemic models greatly rely on proper assumptions. Models with good assumptions are theoretically tractable, generalizing well, interpretable and easy to communicate. However, transmission dynamics models without good assumptions often fail to replicate, fail to generalize, fail to predict outcomes of interest and fail to offer solutions to real-world epidemiology problems [22]. How to discover unknown mechanisms and how to make proper assumptions become critical issues.

Deep neural networks [19, 29], as a universal approximator for unknown mappings [23, 27] and nonlinear operators [32], show incredible effectiveness [41] in pattern recognition. Particularly, deep neural networks can be embedded in differential equations [8, 35] and used to learn unknown mechanisms and solve traditional difficult problems with new perspectives. Data-driven methods proposed in recent years, like symbolic regression, sparse identification of nonlinear dynamical systems [5, 6, 25, 39] and dynamic mode decomposition [5, 15, 45], can find the simplest analytic expressions to describe science, engineering and real-world data. Both deep learning methods and data-driven methods are defensible on their own terms and have generated large, productive scientific literatures; however, both approaches have also been subjected to serious criticism. Deep learning

methods can only output black-box terms which are hard to interpret; data-driven methods need fully observed high-quality data, which is impossible in emerging infectious disease transmission dynamics because of difficulties of tracking the number of suspected and asymptomatic infected individuals. The proposed two-step recovering-explaining framework combines these two areas together and handles the weakness of black-box terms in deep learning methods and the weakness of requiring high-quality observed data in equation-search methods. More explorations and applications of this two-step recovering-explaining method can be found in [42, 43, 52].

Due to the limitation of computation abilities, we have not tested the behavioural change disease transmission dynamics models in different settings. We have not added human behavioural change data such as media reports and search data in Google and Baidu. These data are most meaningful and can give a clear picture on evolution of human behaviour patterns in infectious disease. Moreover, much remains to be done in the future to improve the performance of the two-step recovering-explaining framework. A first step forward may be to discover specific neural network architectures for specific mechanisms in epidemiology models. The second step forward may be to improve the trainability of the two-step framework. How to simultaneously train the recovering and explaining steps needs much more theoretical studies. Coupling transmission dynamics models and deep learning with combining explanation and prediction methods is a promising research field in mathematical epidemiology, and any progress in this field can provide new insights on epidemiological phenomena such as evolution of human mobility pattern, shifting of contact matrix and mutation of variants.

Acknowledgments P.S. was supported by the Postdoctoral Fellowship of York University, Toronto, Canada, China Postdoctoral Science Foundation (No. 2020M683445) and the National Natural Science Foundation of China (NSFC, 12101487(PS)); Y.X. was supported by the National Natural Science Foundation of China (NSFC, 12220101001, 11631012(YX)); J.W. was supported by the Canada Research Chair Program (No. 105588-2011, 230720 (JW)). The authors would like to thank the referees for their kind help in reviewing the paper.

References

1. Baydin, A.G., Pearlmutter, B.A., Radul, A.A., Siskind, J.M.: Automatic differentiation in machine learning: a survey. *J. Mach. Learn. Res.* **18** (2018)
2. Bedson, J., Skrip, L.A., Pedi, D., Abramowitz, S., Carter, S., Jalloh, M.F., Funk, S., Gobat, N., Giles-Vernick, T., Chowell, G., de Almeida, J.R., Elessawi, R., Scarpino, S.V., Hammond, R.A., Briand, S., Epstein, J.M., Hébert-Dufresne, L., Althouse, B.M.: A review and agenda for integrated disease models including social and behavioural factors. *Nat. Hum. Behav.* (2021). <https://doi.org/10.1038/s41562-021-01136-2>
3. Brauner, J.M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A.B., Leech, G., Altman, G., Mikulik, V., Norman, A.J., Monrad, J.T., Besiroglu, T., Ge, H., Hartwick, M.A., Teh, Y.W., Chindelevitch, L., Gal, Y., Kulveit, J.: Inferring the effectiveness of government interventions against COVID-19. *Science* (2020). <https://doi.org/10.1126/science.abd9338>

4. Brownlee, J.: Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. *Machine Learning Mastery* (2020)
5. Brunton, S.L., Kutz, J.N.: *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*, first edn. Cambridge University Press, Cambridge (2019). <https://doi.org/10.1017/9781108380690>
6. Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **113**(15), 3932–3937 (2016)
7. Cao, Y., Li, S., Petzold, L., Serban, R.: Adjoint Sensitivity Analysis for Differential-Algebraic Equations: The Adjoint DAE System and Its Numerical Solution. *SIAM J. Sci. Comput.* **24**(3), 1076–1089 (2003). <https://doi.org/10.1137/S1064827501380630>
8. Chen, T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. In: *NeurIPS* (2018)
9. Chen, X.: Ordinary differential equations for deep learning. Preprint (2019). arXiv:1911.00502
10. Costa, A., Dangovski, R., Dugan, O., Kim, S., Goyal, P., Soljačić, M., Jacobson, J.: Fast Neural Models for Symbolic Regression at Scale (2020)
11. Crammer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Crammer, K., Spergel, D., Ho, S.: Discovering Symbolic Models from Deep Learning with Inductive Biases. *Astro-Ph Physicsphysics Stat* (2020). ArXiv200611287
12. Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Williamson, G., Hassabis, D., Kohli, P.: Advancing mathematics by guiding human intuition with AI. *Nature* **600**(7887), 70–74 (2021). <https://doi.org/10.1038/s41586-021-04086-x>
13. De Brouwer, E., Simm, J., Arany, A., Moreau, Y.: GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. *Cs Stat* (2019). ArXiv190512374
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint (2018). arXiv:1810.04805
15. Dylewsky, D., Tao, M., Kutz, J.N.: Dynamic mode decomposition for multiscale nonlinear physics. *Phys. Rev. E* **99**(6), 063311 (2019). <https://doi.org/10.1103/PhysRevE.99.063311>
16. Funk, S., Gilad, E., Jansen, V.A.A.: Endemic disease, awareness, and local behavioural response. *J. Theor. Biol.* **264**(2), 501–509 (2010). <https://doi.org/10.1016/j.jtbi.2010.02.032>
17. Funk, S., Gilad, E., Watkins, C., Jansen, V.A.A.: The spread of awareness and its impact on epidemic outbreaks. *Proc. Natl. Acad. Sci. USA* **106**(16), 6872–6877 (2009). <https://doi.org/10.1073/pnas.0810762106>
18. Funk, S., Salathé, M., Jansen, V.A.A.: Modelling the influence of human behaviour on the spread of infectious diseases: A review. *J. R. Soc. Interface* **7**(50), 1247–1256 (2010). <https://doi.org/10.1098/rsif.2010.0142>
19. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge (2016)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
21. Hoffman, M.D., Gelman, A.: The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Cs Stat* (2011). ArXiv11114246
22. Hofman, J.M., Watts, D.J., Athey, S., Garip, F., Griffiths, T.L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M.J., Vazire, S., Vesplignani, A., Yarkoni, T.: Integrating explanation and prediction in computational social science. *Nature* **595**(7866), 181–188 (2021). <https://doi.org/10.1038/s41586-021-03659-0>
23. Hornik, K., Stinchcombe, M., White, H.: Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Netw.* **3**(5), 551–560 (1990). [https://doi.org/10.1016/0893-6080\(90\)90005-6](https://doi.org/10.1016/0893-6080(90)90005-6)
24. Jin, Y., Fu, W., Kang, J., Guo, J., Guo, J.: Bayesian Symbolic Regression. *Stat* (2020). ArXiv191008892

25. Kaheman, K., Kutz, J.N., Brunton, S.L.: SINDy-PI: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proc. R. Soc. Math. Phys. Eng. Sci.* **476**(2242), 20200279 (2020). <https://doi.org/10.1098/rspa.2020.0279>
26. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A* **115**(772), 700–721 (1927)
27. Kratsios, A.: The universal approximation property. *Ann. Math. Artif. Intell.* **89**(5–6), 435–469 (2021). <https://doi.org/10.1007/s10472-020-09723-1>
28. Lahoz, B.K.W., Menard, R.: Data Assimilation. Springer, New York (2010)
29. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015). <https://doi.org/10.1038/nature14539>
30. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T.T., Wu, J.T., Gao, G.F., Cowling, B.J., Yang, B., Leung, G.M., Feng, Z.: Early transmission dynamics in Wuhan, China, of Novel Coronavirus–infected pneumonia. *N. Engl. J. Med.* **382**(13), 1199–1207 (2020). <https://doi.org/10.1056/NEJMoa2001316>
31. Liu, R., Wu, J., Zhu, H.: Media/psychological impact on multiple outbreaks of emerging infectious diseases. *Comput. Math. Methods Med.* **8**(3), 153–164 (2007). <https://doi.org/10.1080/17486700701425870>
32. Lu, L., Jin, P., Pang, G., Zhang, Z., Karniadakis, G.E.: Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.* **3**(3), 218–229 (2021). <https://doi.org/10.1038/s42256-021-00302-5>
33. Lu, Y., Zhong, A., Li, Q., Dong, B.: Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In: International Conference on Machine Learning, pp. 3276–3285. PMLR (2018)
34. Niu, M.Y., Horesh, L., Chuang, I.: Recurrent Neural Networks in the Eye of Differential Equations. *Quant-Ph Stat* (2019). ArXiv190412933
35. Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A.J.: Universal differential equations for scientific machine learning. *CoRR abs/2001.04385* (2020). <https://arxiv.org/abs/2001.04385>
36. Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019). <https://doi.org/10.1016/j.jcp.2018.10.045>
37. Rubanova, Y., Chen, R.T.Q., Duvenaud, D.: Latent ODEs for Irregularly-Sampled Time Series. *Cs Stat* (2019). ArXiv190703907
38. Ruthotto, L., Haber, E.: Deep neural networks motivated by partial differential equations. *J. Math. Imag. Vis.* **62**(3), 352–364 (2020). <https://doi.org/10.1007/s10851-019-00903-1>
39. Schaeffer, H.: Learning partial differential equations via data discovery and sparse optimization. *Proc. Math. Phys. Eng. Sci.* **473**(2197), 20160446 (2017). <https://doi.org/10.1098/rspa.2016.0446>
40. Schmidt, M., Lipson, H.: Distilling free-form natural laws from experimental data. *Science* **324**(5923), 81–85 (2009)
41. Sejnowski, T.J.: The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci. USA* **117**(48), 30033–30038 (2020). <https://doi.org/10.1073/pnas.1907373117>
42. Song, P., Xiao, Y.: Estimating time-varying reproduction number by deep learning techniques. *J. Appl. Anal. Comput.* **12**(3), 1077–1089 (2022)
43. Song, P., Xiao, Y., Wu, J.: Methods coupling transmission models and deep learning. Preprint (2022)
44. Sooknanan, J., Mays, N.: Harnessing social media in the modelling of pandemics-challenges and opportunities. *B. Math. Biol.* **83**(5), 57 (2021). <https://doi.org/10.1007/s11538-021-00895-3>

45. Tu, J.H., Rowley, C.W., Luchtenburg, D.M., Brunton, S.L., Kutz, J.N.: On Dynamic Mode Decomposition: Theory and Applications. *J. Comput. Dyn.* **1**(2), 391–421 (2014). <https://doi.org/10.3934/jcd.2014.1.391>
46. Udrescu, S.M., Tegmark, M.: AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.* **6**(16), eaay2631 (2020)
47. Verelst, F., Willem, L., Beutels, P.: Behavioural change models for infectious disease transmission: A systematic review (2010–2015). *J. R. Soc. Interface* **13**(125), 20160820 (2016). <https://doi.org/10.1098/rsif.2016.0820>
48. Weitz, J.S., Park, S.W., Eksin, C., Dushoff, J.: Awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations. *Proc. Natl. Acad. Sci. USA* **117**(51), 32764–32771 (2020). <https://doi.org/10.1073/pnas.2009911117>
49. Xiao, Y., Tang, S., Wu, J.: Media impact switching surface during an infectious disease outbreak. *Sci. Rep.* **5**, 7838 (2015)
50. Yan, Q., Tang, S., Gabriele, S., Wu, J.: Media coverage and hospital notifications: correlation analysis and optimal media impact duration to manage a pandemic. *J. Theoret. Biol.* **390**, 1–13 (2016). <https://doi.org/10.1016/j.jtbi.2015.11.002>
51. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. Preprint (2021). arXiv:2110.11334
52. Yin, S., Wu, J., Song, P.: Optimal control by deep learning techniques and its applications on epidemic models. Preprint (2022)
53. Zhang, L., Tao, Y., Shen, M., Fairley, C.K., Guo, Y.: Can self-imposed prevention measures mitigate the COVID-19 epidemic? *PLoS Med.* **17**(7), e1003240 (2020). <https://doi.org/10.1371/journal.pmed.1003240>
54. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: A fresh approach to numerical computing. *SIAM Rev.* **59**(1), 65–98 (2017)

Chapter 3

Understanding Epidemic Multi-wave Patterns via Machine Learning Clustering and the Epidemic Renormalization Group



Shahram Vatani and Giacomo Cacciapaglia

3.1 Introduction

With the growth of the human population and of the impact on the environment, our societies are becoming increasingly vulnerable to new infectious diseases, in particular those caused by zoonotic viruses [53]. At present, only 3% of the land ecosystem is untouched by any kind of human activity [48]. Furthermore, human-induced climate change is causing relocation of species and also rapid migration of humans. Economic globalization has further increased the mobility of products, merchandise, and people across countries and continents. All these factors play in favor of the jump of viral pathogens from animal species to humans and their rapid diffusion within the world population. The COVID-19 pandemic, exploded in 2020 and caused by the SARS-CoV-2, is the latest example. The COVID-19 pandemic has also dramatically uncovered the unpreparedness of human society to face the threat of a pandemic and its inability to efficiently cope with it during the emergence of epidemiological waves of exponential increase in the number of infections. Henceforth, it is of paramount importance to model mathematically the spread dynamics and to understand the underlying mechanisms in a simple and effective way.

Mathematical modelling of infectious diseases has a long history dating back more than a century [26–29, 40–42, 54–57]. The pioneering SIR model [33] is the first and most famous example of compartmental models, which are based on

S. Vatani (✉)

Galileo Galilei Institute for Theoretical Physics, Firenze, Italy

G. Cacciapaglia

Institut de Physique des Deux Infinis de Lyon (IP2I), IN2P3/CNRS, Villeurbanne, France

Université Claude Bernard, University of Lyon, Lyon, France

e-mail: giacomo.cacciapaglia@in2p3.fr

subdividing the population of an isolated region into various classes. The spread of the disease is then modeled via a set of first-order differential equations in time that describe the flow of individuals between different compartments. These models are deterministic in the sense that the solutions are uniquely determined through their initial conditions, as the flow among compartments depends on fixed rates. These models can be refined by further increasing the number of compartments depending on biological, geographical, and/or social particularities of the case under study. We refer the reader to some excellent reviews available in the literature for more details, e.g., [1, 30, 46, 65, 66]. Other complementary models are, instead, of an essentially stochastic nature. The microscopic processes leading to the spread of the disease are understood in a probabilistic sense and time is typically a discrete variable. Models of this type include lattice and percolation models, many of which are inspired by chemical or particle diffusion processes. We refer the reader to the excellent reviews [22, 61] for more details. These models are related to compartmental ones mentioned above via processes that effectively reduce the number of degrees of freedom, such as mean field approximations and averaging procedures. While very different in their original approaches to the problem, all models feature criticality and symmetries related to a rescaling of time. This similarity to phase transitions in physics has led to further approaches using universality classes of field theories, such as in the pioneering works [13, 18–20, 25, 45].

The connection to critical phenomena has recently led to a more effective incorporation of large and short timescale invariance, as proposed in [16]. Here, the organization of epidemic curves around temporal symmetry principles was dubbed *epidemic Renormalization Group* (eRG) framework [16]. This approach will constitute the core of this brief review. The eRG approach stems from the formal identification of the running of the coupling strength in theories of fundamental interactions [68, 69] with the time dependence of epidemiological quantities, such as the cumulative number of infected individuals [16]. The reorganization around symmetry principles led to the discovery that all episodes of exponential increase in the number of infections, called “waves,” can be described by a universal logistic function, by a solution of the eRG equation, and depending on only two parameters. The simplicity of the approach allows to include human interactions and mobility across different regions of the world [11], leading to an effective prediction of the second wave in Europe in the fall of 2020 [8]. Mobility data can also be included, such as those provided by Google and Apple [7] as well as flight data [5]. The framework has been extended to contain quasi-fixed points [9, 12] to provide a first fully consistent mathematical description of multi-wave pandemics. Last but not least, a slight modification of the approach has proven effective to incorporate the first impact of vaccination campaigns [5] and of health passes [10].

In this chapter, we will briefly review the foundations and properties of the eRG approach, including the mathematical modelling of a single wave and of multi-wave patterns. Then, we will show how a simple machine learning analysis of genome data [15], integrated to the eRG framework, has led to the formulation of a variant-based epidemiological theory [3]. The findings demonstrate that the variant

dynamics is one of the main engines behind the emergence of wave patterns for COVID-19. This result can be used as a template for similar infectious diseases.

3.2 Renormalization Group Epidemiology: From eRG to CeRG

We first review the epidemiological Renormalization Group (eRG) approach [16], based on symmetry principles borrowed and adapted from theoretical physics. This approach introduces an original point of view for the study of the evolution of the number of infected individuals, and it can be combined with the results of the machine learning analysis of the virus mutations, as we will see.

Historically, the first consistent approach to the mathematical study of epidemics was introduced in 1927 by the pioneering work of W.O. Kermack, A.G. McKendrick, and G.T. Walker [33] and goes under the name of compartmental models. This approach is still the most widely used one today. In this approach, the population of a given region is compartmentalized in subpopulations that have different roles in the dynamics of the disease spread. Hence, a set of differential equations is designed to describe the time evolution of the various compartments and the transfer of individuals from one compartment to another. The simplest and most famous case is the SIR model, characterized by three compartments defined by (S)usceptible, (I)nfected, and (R)emoved individuals. The compartment S contains sane individuals that are susceptible to being infected if they encounter an infectious individual; the compartment I consists of infectious individuals; finally, the compartment R contains all individuals that ceased to be infectious and that cannot contract the disease any more for a variety of reasons, from immunization to isolation (quarantine), to death. The SIR model mathematically consists of three coupled differential equations:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\gamma_0 S(t) \frac{I(t)}{N}, \\ \frac{dI(t)}{dt} &= \gamma_0 S(t) \frac{I(t)}{N} - \epsilon_0 I(t), \\ \frac{dR(t)}{dt} &= \epsilon_0 I(t). \end{aligned} \tag{3.1}$$

The parameters γ_0 and ϵ_0 represent constant rates of infection and removal, respectively, while $N = S(t) + I(t) + R(t)$ is the total population considered. If N is constant, only two independent equations remain. This model can effectively describe the diffusion of the disease within an isolated population, and its solutions describe a single epidemic wave. This class of models is deterministic; however they have been shown to emerge from stochastic models like percolation ones [13]. For the interested reader, we point to a pedagogical review in [4]. Compartmental

models have been very successful and are still widely used: as an example, the SIR model allows to define a characteristic number, the reproduction number defined as the ratio $R_0 = \gamma_0/\epsilon_0$, that can tell if the disease is in its exponentially growing phase (for $R_0 > 1$) or declining (for $R_0 < 1$). However, the value of R_0 is usually time-dependent; hence a SIR model with constant parameters cannot correctly describe an epidemic diffusion.

Compartmental models are based on a microscopic description of the diffusion mechanisms of the disease agent and on the specific characteristics of the population. Hence, extensions of this model have been proposed, which add more details on the human behavior and on the characteristics of the individuals with respect to the disease. For instance, one could add vaccinations by including a compartment of vaccinated individuals (SIRV models) [23, 43]. Another direction consists in including exposed individuals, who have contracted the virus but are not infectious yet (SIER model), or other categories. Another direction is age stratification of the models, hence describing the difference in behavior and infection probabilities of various age ranges [2]. In all these models, the increase of compartments increases the number of parameters, the latter being potentially time-dependent. Hence, the predictive power of this approach decreases with the number of compartments, adding more complexity to the system. It is thus natural to ask about other principles to describe pandemics. The eRG approach is special as it goes in the direction of simplifying the description of the disease diffusion, hence reducing the complexity of the mathematical model.

3.2.1 *The Single-Wave eRG Approach*

The eRG offers an alternative approach to study the spread dynamics of epidemics, by focusing on broad and general features of the temporal dynamics instead of the details of the diffusion. In this sense, it is a complementary approach to that described by compartmental models. The eRG is loosely based on the framework of the Wilsonian Renormalization Group (RG) [68, 69], used in physics to characterize physical systems at different scales. The main idea is to identify a quantity that describes the most important properties of the system at different scales of distance or energy. For example, particle collisions can be characterized by the interaction strength, like electromagnetic interactions of electrons and positrons. However colliding them at different energy scales reveals that the strength of the interaction is not the same. The RG framework allows to mathematically describe the flow of the coupling “constant” with energy via a differential equation, called the “beta function.” In analogy to the RG framework, the eRG aims at describing the temporal flow of the disease in terms of a single differential equation. The first step is, therefore, to identify the key quantity that most appropriately describes the status of the disease in a quasi-isolated region.

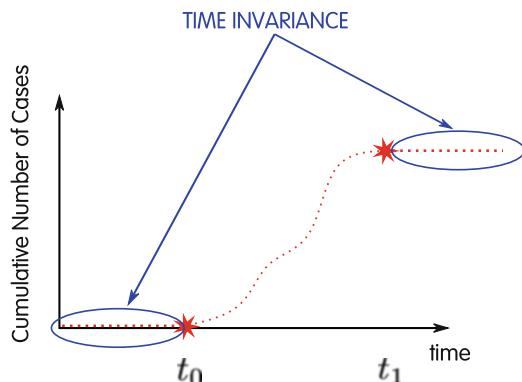
Before presenting the details, a little reasoning can make us appreciate the essence of the eRG by just reinterpreting effectively what happens during the

diffusion of an infectious disease. This argument was the initial motivation for the proposal of the eRG and goes as follows: If an epidemic starts at a time t_0 in a given quasi-isolated population, the cumulative number of infected individual $I_c(t)$ (as a function of time) must be 0 before t_0 and then grows. The function should increase up to a certain limit when it reaches a plateau (around a time t_1). The epidemic is then over and the height of the plateau corresponds to the total number of individuals in the population that have been affected by the infectious disease. Note that, contrary to the compartmental model in Eqs. (3.1), $I_c(t)$ cannot be reduced to a compartment of the population, as it is comprised of all individuals that have been infected. As such, it is different from the number of infectious individuals in Eqs. (3.1). The value of the function at the plateau after t_1 could be equal to the total population, or a value somewhat below depending on the characteristics of the infectious disease. In Fig. 3.1 we show a sketch of such dynamics. Now our primary concerns about $I_c(t)$ are as follows:

- How does it evolve from t_0 to later times?
- Can we determine the height of the plateau?
- Can we determine t_1 ?
- Can a single differential equation with constant parameters describe this time evolution?

From the RG framework point of view, two interesting regions emerge in Fig. 3.1: for $t < t_0$ and for $t > t_1$, the number of infected individuals is constant, implying that $dI_c(t)/dt = 0$. Hence, in these two regions, the system is at “fixed points” of the time evolution flow. These regions present an enhanced symmetry under time translation, which is however broken for times between $t_0 < t < t_1$, where the system flows from the fixed point at $I_c = 0$ to the fixed point at $I_c = I_c(t > t_1)$. How can we describe the flow between these two points? A very simple manner to implement the two fixed points in the flow of $I_c(t)$ is to require that the cumulative number of infected follows this differential equation:

Fig. 3.1 Sketch of the time evolution of the cumulative number of infected individuals $I_c(t)$. The epidemic wave starts at time t_0 , where some infections are introduced, and ends at t_1 , where the number of infected reaches a plateau



$$\frac{dI(t)}{dt} = -\lambda I(t) \left(1 - \frac{I(t)}{A}\right), \quad (3.2)$$

where we have dropped the subscript for convenience. The right-hand side trivially consists of a second-degree polynomial in $I(t)$. Note that if $I(t_*) = 0$ or A at any time t_* , then the solution is a trivial constant. However, as long as $0 < I(t_*) < A$, the solution is a so-called logistic function:

$$I(t) : t \mapsto \frac{A}{1 + e^{-\lambda t + k}}. \quad (3.3)$$

The solution exhibits the characteristic “S” shape seen in many diffusion processes (see Fig. 3.2). The logistic function above, in fact, has been used to describe phenomena that feature the so-called logistic growth, while we obtained it naturally as a solution to the differential equation that came out of our simple reasoning based on fixed points. The parameter A corresponds simply to the value of the function at the plateau, i.e., the final number of cumulative infected, while λ is related to the speed of the spread and effectively has the dimension of a rate. Those are the only two parameters appearing in the differential equation, which need to be fitted on the data. Finally, the parameter k is an integration constant and plays the role of shifting time, i.e., it determines the beginning of the epidemic wave.

Our little reasoning has already made use of the vocabulary from the RG framework, such as “fixed point” and “time invariance.” This motivates the interpretation of Eq. (3.2) as a β -function from the field theory point of view, even though a direct connection has not been established yet (more details about this can be found in [4]). Thus we identify

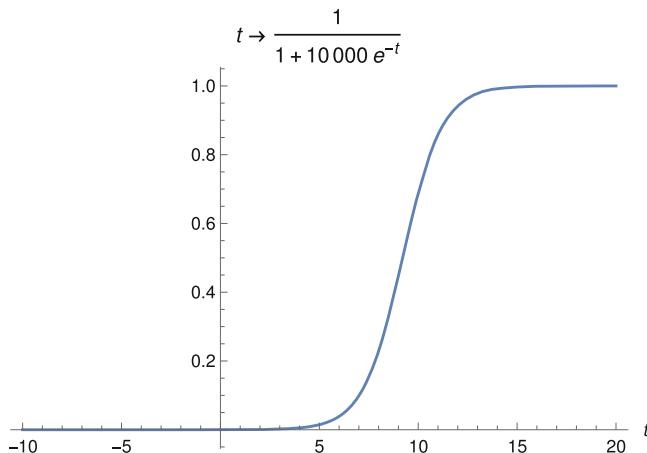


Fig. 3.2 Schematic solution to the eRG equation (3.2) in arbitrary time units

$$\beta(I) = -\frac{dI(t)}{dt} = \lambda I(t) \left(1 - \frac{I(t)}{A}\right). \quad (3.4)$$

Like the RG equations in particle physics, the eRG equation also describes a flow, in time instead of in energy. The interaction strength (like the electric charge) is replaced by an “epidemiological charge,” which is here chosen to be the cumulative number of infected individuals. Like in the RG approach, the diffusion dynamics of the infections is characterized in terms of symmetries: the scale transformations used in the RG in physics are now replaced by the progression in time of the disease diffusion, where fixed points correspond to time-invariant periods. The latter have negligible number of new infections, hence keeping the cumulative number of infections constant in time. Henceforth, the main relation between the RG in physics and the eRG in epidemiology stands on the analogy between symmetries in scale (energy) for the former and symmetries in time for the latter. We summarize a dictionary between the RG and the eRG in Table 3.1.

In practice, however, we can be more general and express the epidemiological charge as a monotonic, differentiable function of the cumulative number of infected, $\mathcal{F}(I)$. A generalized β -function will be given by

$$\beta(I) = -\frac{d\mathcal{F}(I(t))}{dt} = \lambda \mathcal{F}(I) \left(1 - \frac{\mathcal{F}(I)}{\mathcal{F}(A)}\right). \quad (3.5)$$

Following the analogy with the RG framework, different choices of \mathcal{F} correspond to different renormalization schemes. Precisely as it happens in the RG framework, the symmetry-based physical properties do not depend on the scheme: here, this is represented by the fact that the value of the fixed points is preserved. However, the dynamical flow between the two fixed points depends on the scheme choice.

The logistic function obtained as a solution to the eRG equation (3.2) has been used in [16] to describe the diffusion of the first waves of COVID-19 at the beginning of 2020 and also applied to the previous case of SARS. All data show that the logistic function can correctly describe the time evolution of the number of infected individuals in most countries in the world where COVID-19 affected a significant number of people. In fact, it was found that $\mathcal{F}(I) = \ln I$ was better suited in describing the data than I itself. Furthermore, the eRG was extended to include interactions among different quasi-isolated regions in [11], allowing to

Table 3.1 Dictionary between the RG and eRG

RG		eRG
Energy μ	\iff	Time t
Interaction strength (coupling)	\iff	Cumulative number of infected (epidemiological charge)
Fixed points	\iff	Beginning and end of the wave (time invariance)
Scale transformation	\iff	Time progression (diffusion dynamics)

predict the arrival of a second wave in Europe with a few months advance [8]. Further verification has been obtained by describing the epidemiological data in the USA in [6].

The success of the eRG stands in the fact that a single epidemic wave, with its complete exponential growth, can be efficiently described in terms of two constant parameters, the effective infection rate λ and the total number A . In contrast, a SIR model can describe the same dynamics only by allowing time-dependent parameters in Eqs. (3.1), as quantitatively demonstrated in [17]. Hence, SIR-type models can only be used to forecast the development of the disease over short time periods, where the parameters remain approximately constant, while the eRG captures the global evolution of the infections over a time period of an exponential wave. Both models, so far, can only describe a single wave. However, as we will discuss in the next section, the multi-wave patterns can be easily incorporated in the eRG framework, while, for compartmental models, the only remaining option is to introduce it manually via the parameter time dependence.

3.2.2 *The Multi-wave CeRG Approach*

Most models of diffusion of infectious diseases predict that the number of infections will stop increasing after most of the population is infected or immunized. Instead, real diseases, like COVID-19, show that each exponential wave is followed by a period characterized by a nearly constant number of new cases. This is a prelude of a new phase of exponential growth. Schematically, this is shown in Fig. 3.3. Multiple waves can be modeled by introducing time-dependent infection rates or mechanisms that introduce new susceptible populations in the compartmental model. However, the phase in between waves is not well characterized. In the eRG framework, the presence of fixed points would predict a very flat post-wave period, as shown in the left panel of Fig. 3.3.

However, theoretical particle physics suggests a way to consistently model this phase in the eRG approach. In fact, some models of physics beyond the standard model require that a coupling among particles needs to enter a phase where it evolves very slowly with the energy of the process [14, 31, 32]. This entails a near-conformal dynamics, which has approximate invariance under rescaling of energy and length units. Mathematically, this is an indication that the system has approached a fixed point; however some dynamics forbids the system to get too close. This kind of dynamics has been dubbed “walking,” and in analogy we will refer to the epidemiological phase as “strolling” [12]. Mathematically, the appearance of a strolling phase after a wave can be described by moving the second fixed point in the eRG equation to the complex plane. This leads to the following complex eRG (CeRG) differential equation [12]:

$$\beta(I) = -I(t) \left[(1-I)^2 - \delta \right]^p, \quad (3.6)$$

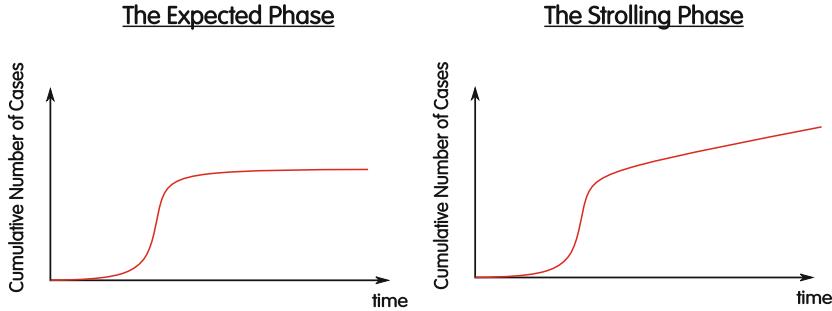


Fig. 3.3 Schematic illustration of an epidemic wave, characterized by an exponential increase in the number of infected individuals. In the left panel, the number of cumulative infections stops at a fixed point, as predicted by the eRG framework, while on the right it enters a phase of constant growth, the strolling phase

where for simplicity we set $\lambda = 1$ and $A = 1$. The model has two additional parameters compared to the eRG: p , which models how fast the system approaches the second fixed point, and δ , which shifts the nontrivial fixed point away from 1. In particular, besides the trivial $I = 0$ one, the CeRG equation has two zeros:

$$I_{1/2}^* = 1 \pm \sqrt{\delta}. \quad (3.7)$$

If $\delta > 0$, this simply shifts the relevant eRG fixed point, and the number of infections will stop growing when reaching $I = I_1^* = 1 - \sqrt{\delta}$. The third fixed point at $1 + \sqrt{\delta}$ is never reached.

However, if $\delta < 0$ is negative, then the fixed points in Eq. (3.7) become complex and conjugate to one another. Of course, physically a complex number of infected does not make sense; hence they cannot be reached by the time evolution of $I(t)$. Yet, their presence alters the evolution. The new behavior can be understood by inspecting Fig. 3.4, where we plot the absolute value of the beta function for a complex I (replaced by α , where $\mathcal{R}(\alpha) \equiv I$). The effect of $\delta < 0$ is to split the eRG fixed point $I = 1$ and move the two in the complex plane. The number of infected, $I(t)$, however is forced to remain real, and it will evolve following the red line: from $I(t_0) \sim 0$ it will roll down the valley, hence implying an exponential growth in the number of infections, similar to the eRG. Once it climbs back toward $I = 1$, it will pass through the saddle point between the two complex fixed points and slow down its evolution. The smaller $|\delta|$, the closer the fixed points to the real axis and the longer the solution will stay near $I = 1$. This period of time when $I(t) \sim 1$ is precisely the strolling phase we observe in data. In Fig. 3.5 we show the profile of the beta function on the real axis for various values of δ and the corresponding solutions. We can clearly see that a correlation exists between the value of δ and the duration of the strolling phase.

This result has important implications for the control of infectious diseases, as it reveals that the time when the next wave will emerge can be correlated to the number

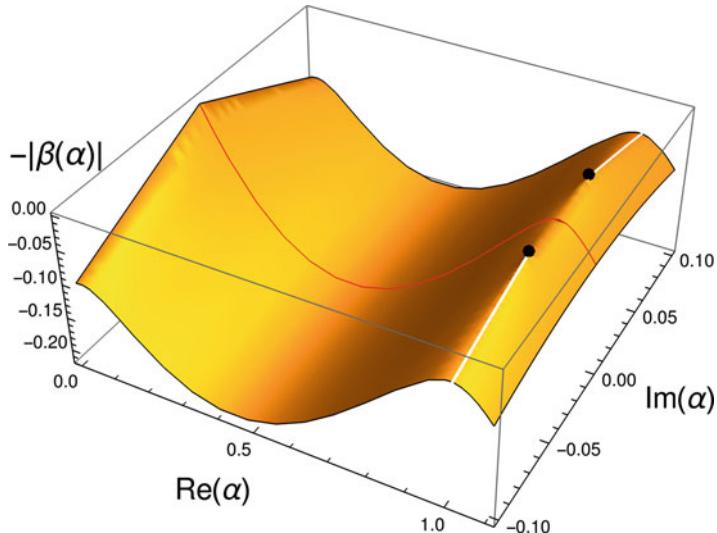


Fig. 3.4 Plot of $-|\beta(\alpha)|$ considering a complex α . The red line represents the trajectory on the real plane, emerging from the real fixed point at $\alpha = 0$ (red curve): the strolling emerges as the solution slows down when passing between the two complex fixed points (black dots)

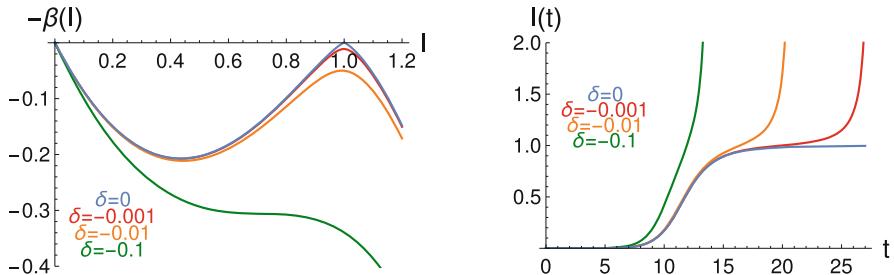


Fig. 3.5 For different values of δ , we plotted on the left the β -function and on the right the corresponding solutions. Both are computed with $p = 0.65$. When δ is too large and negative, the strolling phase disappears

of new infections during the period in between waves. In [12] it was shown how the CeRG model can predict the next wave. We can also deduce that it is important to keep some control measures during the strolling phase, even though the number of new infections may be low [49].

As formulated in Eq. (3.6), the CeRG equation can only describe a single wave with its following strolling phase, as there is no real positive fixed point to stop the growth of the number of infections. This can be easily fixed by adding by hand another real fixed point as follows [9]:

$$\beta(I) = -I(t) \left[(1-I)^2 - \delta \right]^p [1 - \zeta I], \quad (3.8)$$

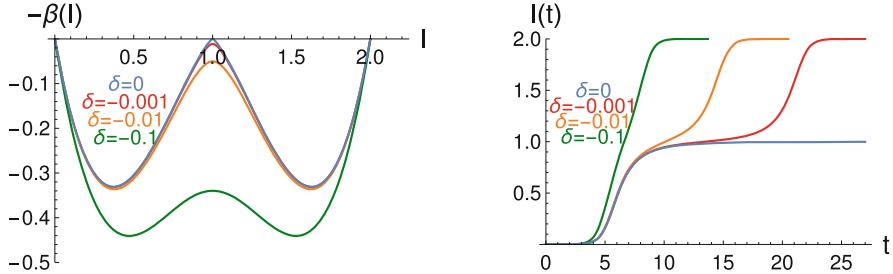


Fig. 3.6 Explicit examples of the beta function in Eq. (3.8) and its solutions with $\zeta = 1/2$. For different values of δ , we plotted on the left the β -function and on the right the corresponding solutions. Both are computed with $p = 0.65$

where the parameter ζ characterizes the height of the second (and last) wave with respect to the first (hence, $\zeta < 1$). Examples of the solutions to this two-wave equation can be seen in Fig. 3.6, while comparison to the data after the second wave can be found in [9].

The CeRG framework can model in a consistent way multi-wave patterns and even give a physical meaning to the inter-wave phase. However, it does not offer any insights on the reason behind the existence of waves. As we will see in the next section, the analysis of virus variants combined with the eRG approach to its dynamics can shed some light on this.

3.3 A Machine Learning Approach to the Wave Pattern

Via phylogenetic analyses of the virus genomes, it has been established that infectious disease agents mutate and generate variants which occur at different times. An example is the seasonal influenza, where each season a different variant is detected and diffuses in the population.

However, in nonseasonal extended pandemics featuring multi-wave patterns, like the “Spanish” influenza of 1918–1919 [62], the diffusion of a virus and the establishment of its variants could not be efficiently approached, mainly due to the paucity of the available data. The COVID-19 pandemic is, on the contrary, revolutionizing our understanding of epidemic diffusion because of the efficient collection of a large amount of data in real time. For example, epidemiological data was crucial in the establishment of the results presented in the previous section. The genome sequencing data has also been made accessible, and it allows for a timely identification of viral variants that successfully radiate throughout the world. As their emergence seems to impact the dynamics of COVID-19 spread, the collection of data available gives a unique opportunity to explore and learn from them. Among the mutations that characterize SARS-CoV-2 variants, those that can be traced along the spike protein (S) sequence are major, although not unique, drivers of viral spread

in the human population for the role that this protein plays in mediating the virus entrance into target cells as well as for its role in mediating escape from antibody responses. We will focus our study of variants on the spike protein. This strategy can be also extended to other types of viruses that feature other surface proteins or different entrance mechanisms.

3.3.1 The Status of Variants

Spike proteins of the same type of virus can differ in their sequence of amino acids. To fix the nomenclature, we define the following:

1. A *mutation* is a single change in the amino acid sequence of the spike protein (substitution, addition, deletion).
2. A *spike variant*, or simply *variant*, is a unique sequence of amino acids in the spike proteins.
3. A *cluster* is a collection of spike variants that are closely related following a given clustering criterion.

Like other coronaviruses, SARS-CoV-2 has relatively low mutation rates [59]; nevertheless the current COVID-19 pandemic has seen the emergence of several epidemiologically relevant variants. Efficient nucleotide sequencing has allowed to track sequence mutations along the genome of SARS-CoV-2 and to identify dangerous variants [35, 47] that appeared to increase the infectivity compared to the initial form that was sequenced from the outbreak in Wuhan, China [70] (GenBank: MN908947.3). Since the second half of 2020, variants of concern (VoCs) and of interest (VoIs) have been identified in various regions of the world, for instance, following the naming scheme of the WHO [34] (Pango lineage [51], GISAID [21, 60]), the Alpha VoC (B.1.1.7, GRY), first identified in September 2020 in the UK [39, 50]; the Beta VoC (B.1.351, GH/501Y.V2) first found in South Africa in May 2020 [63]; the Gamma VoC (P.1, GR/501Y.V3) first detected in Brazil in November 2020 [58], which has been spreading in Manaus notwithstanding the high rate of previous infections; the Delta VoC (B.1.617.2, G/478K.V1) identified in India in October 2020; and the Epsilon VoI (B.1.427+429, GH/452R.V1) found in California in March 2020 [44]. An exhaustive list can be found on the WHO website [67]. These variants have been identified by studying the full genome of the virus, and how new combinations of amino acids correlate with previous ones, without any information from epidemiology.

Other specificities of biological relevance can be obtained by other methods. Considering the Alpha VoC as an example, it has been possible to study its infectious power in lab experiments, finding a higher rate of transmission by 67–75%, compared to the previous ones [39]. The transmission advantage has been confirmed by epidemiological data in the UK [52, 64]. Most analyses of the epidemiological data are done applying the time-honored compartmental models of the SIR type [33, 46, 65], appropriately extended by including more compartments [24]. The

main drawback in this approach, as we already mentioned, is the large number of parameters, which need to be fixed by hand or extracted from the data. Furthermore, the compartments were defined around variants constructed from a biological point of view. It seems natural to ask if this is the right method to understand profoundly the spread dynamics.

For that purpose we used the information directly extracted from the spike protein sequence via a simple machine learning approach. We specifically keep the algorithm untrained and unsupervised, so that our results are not biased by previous knowledge of the variants. In this way, we want to identify and define emerging variants, based uniquely on the sequence of the spike protein. Combined with a simplified and effective approach based on theoretical physics methods, the eRG framework [8, 11, 16], this approach allows us to analyze, at the same time, the variability of the SARS-CoV-2 spike protein in multiple countries and regions of the world and thus provide a direct comparison of the epidemiological impact of the different spike variants.

3.3.2 Method

To start our ML analysis, spike protein sequences have been extracted from the GISAID repository [21, 60] on a country-specific basis, and the date stamp associated to each sequence has been used to obtain a temporal dimension of viral variant appearance. Note that each genome sequence in the GISAID data collection corresponds to the most frequent spike variant occurring in a single infected individual. The general strategy in defining emerging variants can be schematically divided in the following steps [15]:

- (A) Divide the spike sequences in each country or region in temporal bins, whose duration can be varied (1 week or a month).
- (B) Apply a clustering algorithm to define a set of clusters in each time bin.
- (C) Connect consecutive clusters to form temporal chains. The chain link criterion is defined in such a way that chains share similar spike protein variants.
- (D) Define a criterion that allows to tag some chains as *emerging variants*, while others are dismissed and epidemiologically irrelevant.

Below we describe in detail the algorithm leading to our identification of variants.

3.3.2.1 Cluster Algorithm

In order to cluster spike protein sequences, we need to define a method to compare them. Our choice goes for the Levenshtein measure (LM) [37, 38], which acts on each pair of sequences. This effectively counts the minimal number of amino acid substitutions, deletions, and insertions (i.e., mutations) needed to transform one sequence into the other, and vice versa. This is a naive and unweighted approach,

where all the abovementioned differences are counted effectively by 1. It is worth mentioning that when it comes to sequence differences (in particular for DNA comparison), there are ways to score them according to some biological principles.

Once the LM scores are assigned to each pair of sequences, the algorithm constructs a “tree of proximity” by pairing sequences that are the closest to each other, a process done step by step. Initially all the sequences are considered to be single “leaves”; they are not linked to each other by branches. For n samples, we start thus with n leaves that are like n branches containing only one leaf. Each step will then consist in regrouping two branches together forming a new branch. Naturally there can be at most $n - 1$ of those steps. The two branches grouped have to be the closest and this quality depends on how we want to measure the distance between branches. The LM only compares two individual sequences, but it can be used to define a distance between branches, where different methods are available. To explain how this can be done, let us fix some notation. We will define by $d(s_1, s_2)$ the LM between two sequences s_1 and s_2 , and we will call the distance between two branches $\text{dis}(A, B)$ for branches A and B . The measure $\text{dis}(A, B)$ can be developed in terms of $d(s_1, s_2)$ for s_1 and s_2 being sequences from A and/or B . The most common choices to compute $\text{dis}(A, B)$ are as follows:

- (a) Single Linkage Clustering:

$$\text{dis}(A, B) = \min_{s_1 \in A, s_2 \in B} d(s_1, s_2), \quad (3.9)$$

hence based on the minimal LM among the sequences in the two branches.

- (b) Complete Linkage Clustering:

$$\text{dis}(A, B) = \max_{s_1 \in A, s_2 \in B} d(s_1, s_2), \quad (3.10)$$

based on the maximal LM.

- (c) Unweighted Average Linkage Clustering:

$$\text{dis}(A, B) = \frac{1}{|A||B|} \sum_{s_1 \in A, s_2 \in B} d(s_1, s_2), \quad (3.11)$$

where $|X|$ measures the number of sequences in the branch X .

- (d) Ward Method:

$$\begin{aligned} \text{dis}(A, B) = & \frac{|A| \times |B|}{|A| + |B|} \left[\frac{\sum_{s_1 \in A, s_2 \in B} d(s_1, s_2)^2}{|A| \times |B|} \right. \\ & \left. - \frac{\sum_{s_1 \in A, s_2 \in A} d(s_1, s_2)^2}{2|A|^2} - \frac{\sum_{s_1 \in B, s_2 \in B} d(s_1, s_2)^2}{2|B|^2} \right]. \end{aligned} \quad (3.12)$$

An intuitive interpretation of the Ward method can be obtained if the data are living in a Euclidean space. Then, considering the center of gravity of each branches $g_{A/B}$, the formula is simplified to

$$\text{dis}(A, B) = \frac{|A| \times |B|}{|A| + |B|} d(g_A, g_B), \quad (3.13)$$

i.e., it measures the distance between the centers of gravity of the two branches. However, our sequence data measured via the LM do not live on a Euclidean space.

Note that (a) and (b) do not depend on the multiplicity of each unique spike variant in the branch, while (c) and (d) do.

The construction of the tree of proximity is a bottom-up approach. The tree is completed when all sequences are grouped into a single branch (at the step $n - 1$), as illustrated in Fig. 3.7. In [15], the Ward method was chosen, being very popular among the different hierarchical clustering algorithms; however we have checked that other methods lead to similar result. The final step in the clustering algorithm consists in defining the clusters. This is done via a *cutoff* in the branch distance so that branches whose Ward distance is larger than the cutoff are considered as separate clusters. We applied the same cutoff to all branches. In Fig. 3.7 we illustrate this process with an example of such a tree built on real data, where the five colors correspond to the five different branches cut by the cutoff. Each one, therefore, defines a cluster. Note that the value of the cutoff can be chosen by the

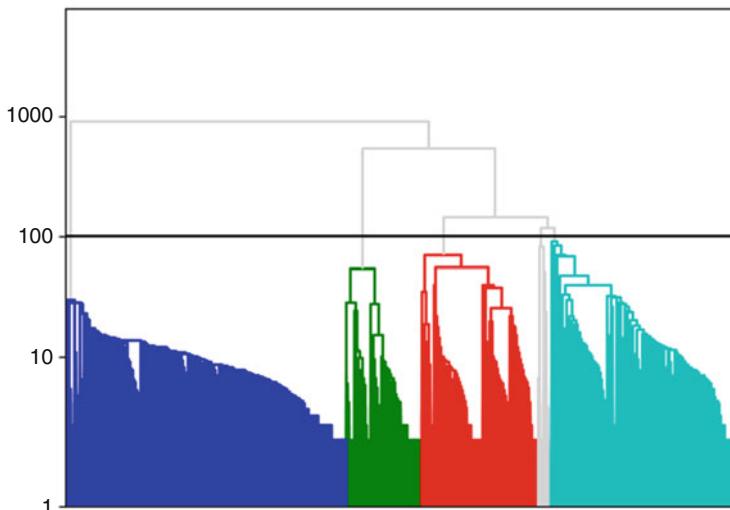


Fig. 3.7 Example of proximity tree built on real sequence data. The y-axes report the value of the Ward distance, with a cut at distance 100 defining 4 major clusters, in colors. Branches that pass the cut but contain less than a 1% threshold of the sequences in the initial sample are removed

data themselves, as the number of clusters is roughly stable for a given range of cutoff values [15]. Furthermore, many small clusters also appear, containing very few sequences: this could be due mainly by the presence of sequencing errors. To reduce the noise, one can impose a *threshold* in the percentage of sequences contained in each cluster and remove clusters that do not pass the threshold.

3.3.2.2 Emerging Variants as Persistent Time-Ordered Cluster Chains

The time evolution and emergence of SARS-CoV-2 variants can be studied by applying our precedent clustering algorithm to the spike sequence data binned in time, by calendar month or weeks or any other appropriate length. After clustering the data in each bin, following the procedure described above, the clusters in consecutive time bins are compared. The aim is at defining links between them. Chains are based on *strong links* defined between clusters whose dominant variant is the same. As an illustration, we schematically show the outcome of this procedure in Fig. 3.8, where strong links are indicated by thick black lines. Furthermore, clusters that do not have any strong links with the previous time bin are linked via *branching links* to the precedent cluster that have the closer similarity with it. This similarity is defined in terms of the LM of the spike variant in the two clusters. The branching links allow to define connections between new chains/clusters and existing ones, and they are indicated by thin gray lines in Fig. 3.8.

Our definition of *emerging variant* is that of a chain that lasts more than a threshold of time bins, which we establish from data to be 3. Looking at Fig. 3.8 as a reference, the chain **1–4–7–10** is an emerging variant. On the other hand, **2–5–8** does not pass the threshold as it does not continue more than 3 consecutive time

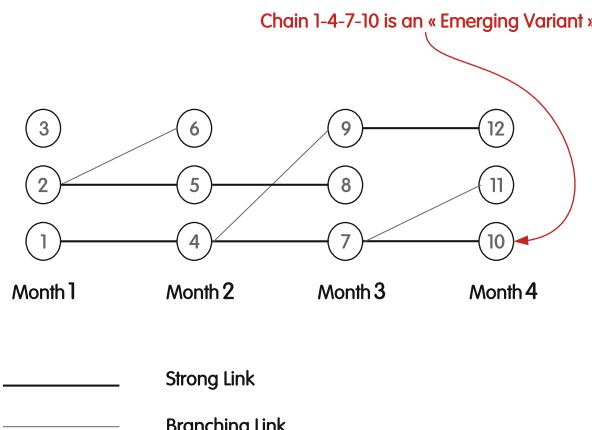


Fig. 3.8 Schematic representation of the chain reconstruction from the machine learning algorithm. Chains connected by strong links and spanning more than 3 time bins are identified with emerging variants

bins. We also see the emergence of a potential new variant in the chain **9–12**, which is connected to the other emerging variant by the branching link between **4** and **9**. Further data would be needed to establish the fate of this chain.

For clarity, we summarize the definitions used in defining chains:

1. A *dominant variant* in a cluster is the spike variant that is most frequently appearing in the cluster. Note that the chains are created by linking consecutive clusters when they possess the same dominant variant. Subdominant variants in each cluster are retained if their frequency is above 1% in the cluster.
2. An *emerging variant* is defined as an established chain that contain more than three consecutive clusters, defined using our linkage algorithm. This criterion was established empirically from the results of the chain reconstruction.

It should be noted that some of the emerging variants defined by our procedure can be associated to VoCs and VoIs, as defined by the WHO, as they share the same characteristic spike mutations.

3.3.3 Application to COVID-19 Data

The method described above has been applied to real COVID-19 data in [15]. Besides a novel and unbiased definition of emerging variants, one can associate each variant to the number of infections and correlate the emergence of variants to that of epidemiological waves.

As England has the largest sequencing sample available on GISAID, with 646.697 sequences as of the end of August 2021, the England dataset was chosen as the main focus in the analysis. Yet, the analysis can be straightforwardly repeated on any geographical region, as long as a sufficient number of sequences are available. This minimizes statistical and sampling bias errors. The sequences are first pruned in order to eliminate those that have transcription errors (i.e., an “X” appearing in the place of an amino acid); henceforth 461.122 sequences were retained, out of which 13.887 distinct ones are identified. As a second step, the sequences are time binned following the date tag in the GISAID repository. As here we want to focus on the wave patterns, a time bin length of 1 month is chosen, which allows us to track the evolution of each variant during the development of a wave, which typically lasts 3 months for COVID-19. Hence, for each month, we run the clustering algorithm on the pruned data to define clusters, retaining only the ones comprising at least 1% of the monthly dataset. The cutoff on the Ward distance r_W between branches, as well as the 1% threshold above, was chosen to optimize the coverage of the dataset (i.e., it is required that the defined clusters cover at least 90% of the data) while keeping the number of clusters below 10.

The results are shown in Fig. 3.9 from [15] for two choices of the Ward distance: $r_W = 100$ in the left and $r_W = 200$ in the right plots. For the two choices, we identified six and four cluster chains, respectively, that last more than 3 months. In the middle and bottom rows of Fig. 3.9, we show the new monthly infections (per

Monthly ML classification of emerging variants

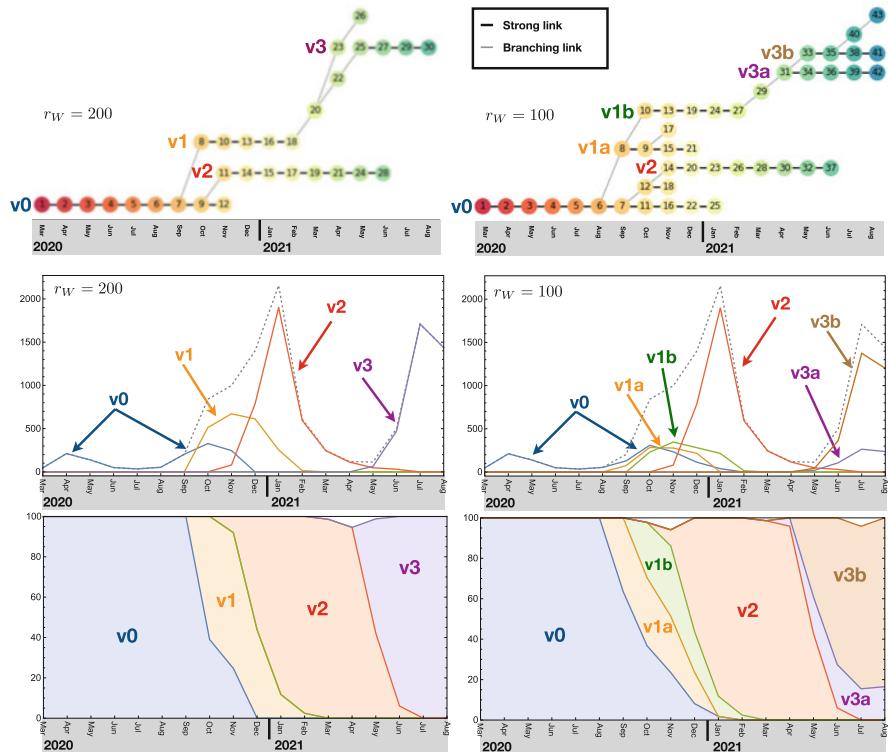


Fig. 3.9 From [15]: application of the clustering algorithm to the sequencing data for England from March 2020 to August 2021. In the top row, we show how clusters are linked to form chains, which are then identified with emerging variants. In the middle and bottom plots, we show the number of monthly infected per 100k inhabitants and percentage of occurrence for each emerging variant. The left plots correspond to a cutoff in the Ward distance of $r_W = 100$ while the right ones to $r_W = 200$. Note that the chains v2 and v3 for $r_W = 200$ can be associated to the Alpha and Delta VoC, respectively

100k inhabitants) and the frequencies of the cluster chains, which we identify as emerging variants in the following: Given that only a small fraction of the infected individuals have their viral charge sampled and sequenced, we estimated the number of people infected by each variant by multiplying the number of positive tests by the rate of occurrence of each variant in the sequencing data, i.e., their frequency. This rough approximation allows us to reliably extract the temporal evolution of each variant in the population.

The results for $r_W = 200$ can be directly compared to the VoCs identified by the WHO: Comparing the frequencies of occurrence, we see that v2 can be associated to the Alpha VoC, while v3 matches the epidemiological data for the Delta VoC. We also checked that the dominant spike variant for the two chains presents the

mutation characteristic of the two VoCs: N501Y, D614Y, and P681H for the Alpha VoC and L452R, T478K, D614G, and P681R for the Delta VoC.

From the case $r_W = 200$, we clearly see that v1, which is responsible for the second wave, branched off from v0 in October 2020. Similarly, v2, which corresponds to the Alpha VoC, also branched off from v0 a month later. The Delta VoC v3, instead, developed from v1 from February to May 2021, via two intermediate clusters, 20 and 22. Finally we see the emergence of a branch, 20–23–26, which died off being dominated by the Delta VoC starting with cluster 25. By lowering the cutoff that defines clusters (see left plots in Fig. 3.9 for $r_W = 100$), one can see how v1 splits in two distinct, but closely related, chains, as well as the Delta VoC v3. The Delta VoC is now seen as branching off from v1b. The closeness of the clusters splitting from v1 and v3 is confirmed by comparing the dominant spike sequences, showing that v1b differs from v1a only by the mutation L18F, while v3b differs from v3a only by the mutation T95I. In particular, cluster 43 emerged in August 2021, and its dominant variant bears the Y145H and A222V mutations that identify the AY.4.2 lineage (“Delta plus” variant) [36], which has been classified by Pango [51] at the beginning of September. Out of the many new lineages that have been recently isolated, only this one is highlighted by the analysis. As such, and with the caveat that our analysis includes only data up to August 2021, the ability of this novel variant to give raise to a stable chain in the near future deserves close attention.

These results firstly show that the phylogenetic relation between variants emerges from this simple algorithm applied exclusively to the spike protein sequence. Furthermore, we see a distinctive pattern relating the emergence of a persistent variant and the exponential increase in infections that ignites a new pandemic wave. A new wave only emerges when a new variant is generated, which has the virological strength to overcome the old ones. This is seen very clearly with v2 (or Alpha VoC) which spins off from v0 closely to v1 and takes over by generating a third wave. We also see the emergence of short-lived variants that do not have the power to start a new wave and therefore die off without infecting a sizable number of individuals. All short-lived chains have less than two clusters; hence we define a minimum length of three for persisting chains.

While v2 and v3 were identified as VoCs by other types of analyses on the genome, our unbiased analysis also uncovers that the second wave, which took place in the fall of 2020, is also related to a new variant stemming out of the original strain. This emerging variant, v1, is very similar to the original v0; however the data suggests that it is its presence that drives the exponential increase during the second wave.

Given the importance of emerging variant in the development of pandemic multi-wave patterns, it is important to also develop an epidemiological model for the variants. This is the main focus of the next section, where we will return to the eRG framework.

3.4 An Epidemiological Theory of Variants: The MeRG Framework

The results of the machine learning analysis strongly suggest that there is a tight relation between the genesis of a new emerging variant and the onset of a new wave, with exponential increase in the number of infections, in the epidemiological data. In [3], based on this observation, we developed a framework that can be used to describe the evolution of each variant. The model is based on the eRG approach by including mutations (MeRG).

3.4.1 The Model

The eRG approach consists in defining β -functions that govern the time evolution of the system at the global level. Here we want to apply the same framework to the various emerging variants, observed in the analysis of the virus genome. To this end, for each variant we assign a cumulative number of infected I_i , with $i = 1, \dots, N_v$. Then we define a set of independent epidemic charges

$$\alpha_i = f_i(I_1, \dots, I_{N_v}), \quad (3.14)$$

where different choices of f_i correspond to different renormalization schemes and, as previously noted, we expect physical results in general not to depend on the choice. This is a generalization of the eRG for a single variant. The β -functions are then defined as

$$-\beta_i = \frac{d\alpha_i}{dt} = \sum_{j=1}^{N_v} \frac{df_i}{dI_j} \frac{dI_j}{dt}(t), \quad \forall i = 1, \dots, N_v. \quad (3.15)$$

The dynamics of the system now depends on two elements: the scheme choice, encoded by the functions f_i , and the evolution of each variant, encoded in the expression for dI_j/dt , i.e., the beta function for each individual variant. In principle, the latter may depend on the presence of other variants.

For the scheme dependence, the simplest choice is to fix $f_i = I_i$. This choice marries well with the fact that the beta function for each variant I_i is a polynomial in I_i and it does not depend on the presence of other variants. This feature is suggested by two analyses:

- In [3] the interdependence of various variants was studied by means of a compartmental model. It was shown that, if the two variants do not differ too much in terms of reproduction number, the two evolve independently to each other, approximately.

- In [15] it was shown that the epidemiological evolution of each variant can be well described by a logistic function, stemming from an eRG equation.

Hence, the variant system can be described by a set of independent eRG equations, one for each variant (MeRG):

$$-\beta_i = \frac{d\alpha_i}{dt} = I_i \lambda_i \left(1 - \frac{I_i}{NA_i} \right). \quad (3.16)$$

In practice, this means that we are modelling the time evolution of each variant as independent of the others. We shall see that this assumption also leads to reasonable results compared to real-world data. Indeed, the system (3.16) allows for an analytic solution of $I_i(t)$ which can be written in terms of logistic functions

$$I_i(t) = \frac{NA_i}{1 + e^{-\lambda_i(t-\kappa_i)}} \quad (3.17)$$

that, as we saw, reproduce very well the data for each wave of the COVID-19 pandemic.

3.4.2 Flow Among Variants: Fixed Points and (Ir)relevant Operators

To obtain a mathematical insight on the variant dynamics, we will restrict the analysis to two variants. The vector field $(-\beta_1(I_i), -\beta_2(I_i))$ in the (I_1, I_2) -plane is schematically plotted in Fig. 3.10 from [3]. It has four fixed points, i.e., points $(I_1^{(k)}, I_2^{(k)})$ (for $k = 0, 1, 2, 3$) where

$$\beta_i(I_1^{(k)}, I_2^{(k)}) = 0, \quad \forall i = 1, 2, \quad (3.18)$$

explicitly given by

$$\begin{aligned} P_0 &= (I_1^{(0)}, I_2^{(0)}) = (0, 0), & P_1 &= (I_1^{(1)}, I_2^{(1)}) = (NA_1, 0), \\ P_2 &= (I_1^{(2)}, I_2^{(2)}) = (0, NA_1), & P_3 &= (I_1^{(3)}, I_2^{(3)}) = (NA_1, NA_2). \end{aligned} \quad (3.19)$$

Among them, P_0 is repulsive in all direction (i.e., in Fig. 3.10 all arrows point away from it) and corresponds to the case where no disease is present. In fact, moving away from this fixed point by infecting even only a small number of individuals of the population (with either of the two variants) causes the system to flow to one of the other three fixed points. Among them, $P_{1,2}$ are repulsive in one direction but attractive in the other. Since they are characterized by $I_2 = 0$ or $I_1 = 0$, respectively, they correspond to the endpoints of scenarios in which

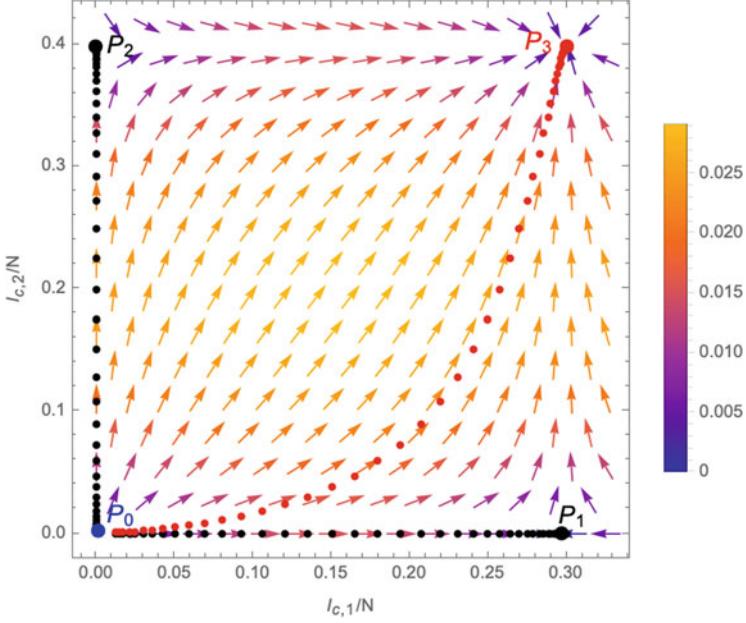


Fig. 3.10 From [3]: schematic structure of the MeRG equations in the two-dimensional plane of the variants (the values used for the plot are $N\lambda_1 = 0.2$, $N\lambda_2 = 0.25$, and $A_1 = 0.3$, $A_2 = 0.4$). The β -functions exhibit four fixed points: $P_{0,1,2,3}$ of which P_0 is repulsive in all directions, $P_{1,2}$ have one attractive direction and one repulsive one, while P_3 is attractive in all directions. Different trajectories as solutions of Eq. (3.16) connecting the fixed points are indicated by the dotted lines. The coloring of the vectors indicates the norm $\|(\beta_1, \beta_2)\|$, which, respectively, leads to smaller or larger distances between different points of the flow lines

variant 2 or variant 1 is never present in the population. These fixed points can be reached only by the flow lines represented in black in Fig. 3.10, which are initiated by a deformation away from P_0 along I_1 or I_2 only, respectively. Any deformation that switches on both I_1 and I_2 (i.e., any scenario in which infected with both variants are present in the population) causes the system to flow to fixed point P_3 ; an example of such a flow is indicated in red in Fig. 3.10. Which trajectory is realized depends on the initial deformation, which is represented by the parameters $\kappa_{1,2}$ in the solution (3.17).

The assumption of a deformation away from the fixed point P_0 in Fig. 3.10 along a generic direction is not realistic, since it would correspond to the simultaneous appearance of infected individuals of both variants in the population. A more likely scenario would be the appearance of one variant first, while a second deformation at a latter stage introduces the second variant. This dynamics can be understood from an RG perspective as a relevant operator being switched on.

To make this statement more precise, we first need to introduce the concept of *critical surface* associated with a fixed point of the β -functions. A critical surface

consists in all points in the (I_1, I_2) -plane from where the RG flow leads to the fixed point in question. Concretely, for the fixed points $P_{1,2,3}$, the critical surfaces are

$$\begin{aligned} C_{P_1} &= \{(I_1, I_2) \in \mathbb{P} | I_1 > 0 \text{ and } I_2 = 0\}, \\ C_{P_2} &= \{(I_1, I_2) \in \mathbb{P} | I_1 = 0 \text{ and } I_2 > 0\}, \\ C_{P_3} &= \{(I_1, I_2) \in \mathbb{P} | I_1 > 0 \text{ and } I_2 > 0\}. \end{aligned} \quad (3.20)$$

A *relevant operator* (from the perspective of the fixed point in question) corresponds to a direction that drives the theory away from the critical surface, such that it flows to a new critical point. In the case at hand, $C_{P_{1,2}}$ have one critical direction orthogonal to it, which, from an epidemiological perspective, precisely corresponds to the appearance of the second variant. A small deformation at any point of $C_{P_{1,2}}$ (e.g., due to a relevant mutation of the virus) causes the system to deviate from the critical surface and ultimately flow toward P_3 .

In a scenario with only two variants, the fixed point P_3 has no relevant deformations and is attractive along all directions. Instead, small fluctuations along trajectories leading toward P_3 (such as the red path shown in Fig. 3.10) can be interpreted as *irrelevant operators* being switched on, as they will not change significantly the flow of the system.

3.4.3 Connecting Variant Dynamics to the CeRG

It was shown in a previous section that multi-wave patterns can be mathematically described in terms of complex fixed points in the CeRG framework. How can the variant dynamics, i.e., MeRG, be related to complex fixed points?

To this end, in [3], it was considered in more detail a particular case of a flow along the critical surface C_{P_1} with a relevant operator being switched on along the way (i.e., the second variant appearing at some moment $t_0 > 0$) through a small fluctuation. The latter drives the system from the proximity of the fixed point P_1 to the new one P_3 . From the perspective of the new fixed point, the second phase of the flow looks like an RG flow from a UV fixed point P_1 to an IR one P_3 .

Since during the first part of the flow, the number of active infected of the second variant $I_2(t)$ is fairly small, this flow is very well approximated by a usual eRG dynamics. Once the system reaches the vicinity of the fixed point P_1 , it will then enter into a quasi-linear growth phase (the strolling phase), in which the number of active infected with respect to both variants is small and therefore the total number of infected $(I_1 + I_2)(t)$ only grows linearly (see Fig. 3.3). However, after a certain time, the number of infected I_2 will grow exponentially (while the number of infected with respect to the original variant remains small), and the system enters into the crossover phase. Now, the β -function for I_2 is, once again, essentially modeled by a standard eRG equation (see Eq. 3.16) describing the flow to P_3 .

In this picture, the two-wave structure is explained as the (more or less successive) appearance of two different variants of the disease. In particular, the strolling phase is explained by the fact that the system comes close to a fixed point, which, however, it cannot reach. It nevertheless spends significant time in its proximity. As we saw earlier, a similar reasoning underlies the CeRG approach modelling the combined number of cumulative infected $I_{\text{tot}} = I_1 + I_2$:

$$-\beta_{\text{CeRG}} = \frac{dI_{\text{tot}}}{dt}(t) = \lambda I_{\text{tot}}(t) \left[\left(1 - \zeta \frac{I_{\text{tot}}(t)}{A} \right)^2 - \delta \right]^{p_1} \left(1 - \frac{I_{\text{tot}}(t)}{A} \right)^{p_2}, \quad (3.21)$$

with A the asymptotic number of infected, λ the infection rate, $\zeta > 1$, and $\delta < 0$. Indeed, besides the fixed points $I_{\text{tot}} = 0$ and $I_{\text{tot}} = A$, the beta function (3.21) also has the complex fixed points $I_{\text{tot}} = \frac{A}{\zeta} (1 \pm i\sqrt{|\delta|})$, which cannot be reached by the flow, but are responsible for the linear growth phase.

To compare the independent approach (3.16) with the β -function in (3.21), we assume that the mutation occurs significantly after the maximum number of infected of the first variant. Furthermore, we can also compare them to the following combined β -function for the total cumulative number of infected:

$$\begin{aligned} -\beta_{\text{tot}} = & \frac{1}{N} \left[\frac{dI_1}{dt}(t) + \frac{dI_2}{dt}(t) \right] \sim \theta((NA_1 - I_1)) \lambda_1 I_1 \left(1 - \frac{I_1}{A_1} \right) \\ & + \theta((I_2 - \xi)(N(A_1 + A_2) - I_2)) \lambda_2 (I_2 - N A_1) \left(1 - \frac{I_2 - NA_1}{NA_2} \right), \end{aligned} \quad (3.22)$$

where θ is the Heaviside step function. It simulates the impact of a relevant operator being switched on. Furthermore, we can use the solutions of $(I_1, I_2)(t)$ as logistic functions to schematically plot (3.22): The top panel of Fig. 3.11 shows a parametric plot of $(I_1(t) + I_2(t), \frac{dI_1 + I_2}{dt}(t))$ for different values of t . The latter are very well approximated by (3.22) (shown by the thin black line), except for a small region around $I_t \sim A_1$, in which the beta function does not in fact reach zero, but interpolates between the two terms in (3.22). This region corresponds to a nontrivial interaction between the variants and governs the transition from the first part of the flow (close to the original critical surface) to the crossover flow. It precisely corresponds to the linear growth region in the context of the CeRG: Indeed, a similar shape of the beta function can also be achieved through a function of the form (3.21), as is shown in the bottom panel of Fig. 3.11. The region around $I_{\text{tot}} \sim A_1$ corresponds to the RG flow not quite reaching a zero, thus leading to the quasi-linear growth phase.

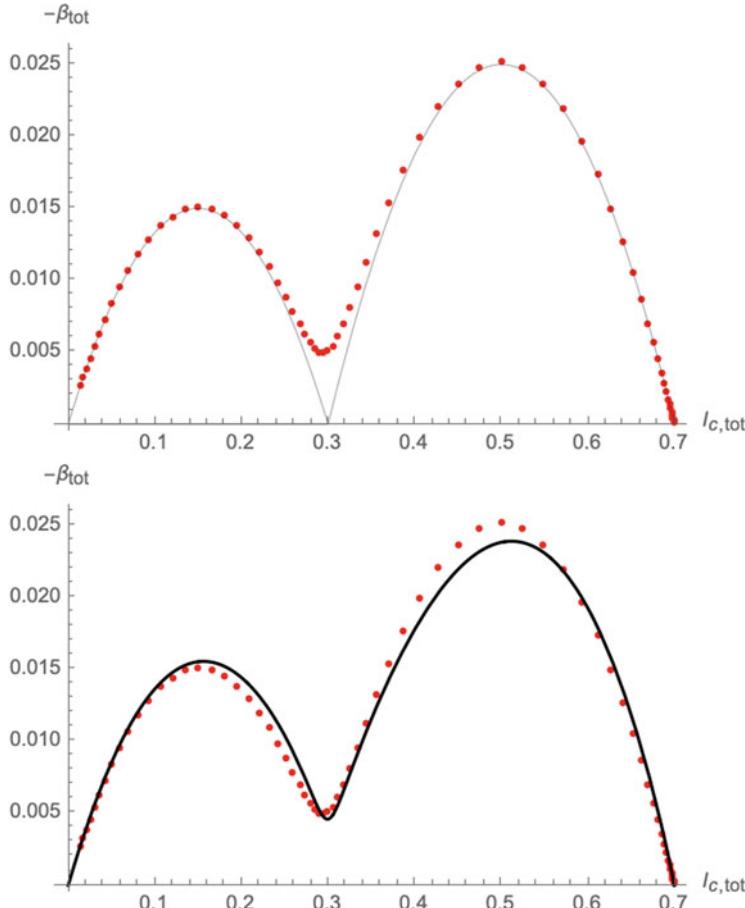


Fig. 3.11 From [3]: schematic plot of the beta function of the total number of infected computed from the solutions (3.17) (red dots). Left panel, comparison to (3.22); right panel, fitting with a β -function of the type (3.8), with $\kappa = 0.2075$, $\delta = -0.0021$, $p_1 = 0.336$, $p_2 = 0.959$, $A = 0.6996$, and $\frac{A}{\zeta} = 0.2995$

3.4.4 Fitting the Real Data

To further validate the epidemiological theory of variants, we need to compare it to real-world data [3, 15].

The MeRG framework models the time evolution of the cumulative number of infected by each variant in terms of an independent logistic function associated to each variant. Hence, for each variant, one needs to fit the infection rate λ_i (in inverse days) and A_i being the total number of affected individuals at the end of the wave (per 100,000 inhabitants). Interestingly, λ_i encode the effective diffusion

speed of each variant, including not only its intrinsic viral power but also the effect of pharmaceutical measures (like vaccinations) and social distancing measures. Hence, if the diffusion occurs under similar social conditions, comparing the λ_i could measure the ability of the new variant to spread and infect new individuals.

We used the eRG logistic function to fit the epidemiological data in various regions of the world, after distributing the new daily infected to each variant proportionally to the variant frequency observed in the sequencing data. For this purpose, we used the full dataset from GISAID, using the VoC classification embedded in the GISAID data. The results are shown in Fig. 3.12, where we show

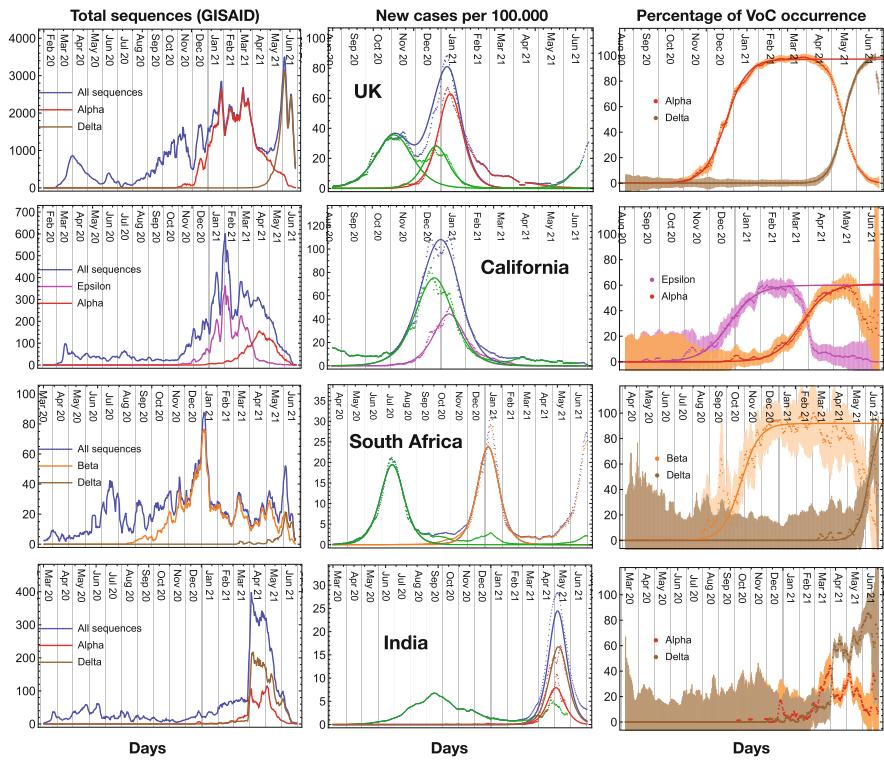


Fig. 3.12 From [3], results of the MeRG fitting of the number of infected associated to each relevant variant. Each row corresponds to a geographical region. In the left column, we show the total number of sequencing available on GISAID (in colors the ones associated to the relevant VoC or VoI); the middle column shows the number of new daily infected (per 100,000 inhabitants); the right column shows the percentage of each VoC or VoI in the sequencing data. All plots show daily rates, with data smoothed over a period of 7 days. In the middle plots, the data are shown by dots, where blue corresponds to the total and the other colors show the number of infected associated to each variant. The solid lines show the result of the fits to the MeRG model (note that only for the UK we fit the “standard variant”—in green—with two logistic functions). In the left plots, the error derives from the expected statistical variation on the number of daily sequences (after smoothing). For all the plots, the classification in variants derived from the GISAID data

the number of sequences (left plot), the new number of infections per variant and the result of the MeRG fit (middle), and the frequency of the VoCs (right). Note that the total numbers are plotted in blue while the VoCs in colors. We considered the epidemiological data from the most recent waves, which developed between September 2020 and February 2021. The analysis is shown for the whole UK, California, South Africa, and India.

For the UK, the green curve in the middle plot shows that, after the first peak at the beginning of November, a second smaller peak developed. We know that this is due to a new variant, v1 in the machine learning analysis; hence we describe the two with two independent functions. The second one is subtracted from the data when fitting for the Alpha VoC data. Similar procedure is applied to all countries. In all cases, the logistic function offers an excellent fit for the diffusion of each variant, hence supporting the MeRG model.

3.5 Conclusion

In this chapter, we have reviewed a new approach to epidemiology, driven by theoretical physics. The main idea is to study the diffusion of an infectious disease based on symmetry reasons, i.e., approximate time-invariant symmetries. This led to the development of the eRG framework, where the evolution of each epidemiological wave, characterized by an exponential increase in the number of infected individuals, can be described in terms of a single differential equation with two constant parameters. This leads to a simple and economical model, showing that each wave is described by a family of logistic functions which have the same basic shape. The model can be adapted to study the diffusion between different quasi-isolated regions.

The main breakthrough that this approach has led to is in the modelling and understanding of multi-wave patterns, observed in many diseases from the Spanish Influenza of 1918–1919 to the COVID-19 pandemic of 2019. In the eRG framework, multi-wave dynamics can be modeled by introducing complex fixed points in the differential equation. This allows to describe consecutive waves, and also the inter-wave period characterized by a relatively small and constant number of new infections. CeRG is the model that described multi-waves with a single differential equation.

Variants stemming from mutations of the virus genome played a key role in the COVID-19 pandemic. Hence, we developed an unbiased and unsupervised machine learning algorithm to define variants uniquely based on the genome of the spike protein, which is mainly responsible for the cellular penetration and for the antigenic response to the infection. Besides matching with the WHO-identified variants, this analysis helped showing that each new wave can be uniquely correlated to the genesis and diffusion of a new variant. Furthermore, the number of infections that can be associated to each variant can be fitted by an eRG logistic function. This leads to the formulation of the MeRG, where each variant is associated to an independent

eRG equation with two relevant parameters. The epidemiological time evolution can now be thought of as a flow in a multidimensional space of variants, where the system flows among unstable fixed point, similar to what we described in the CeRG approach.

The eRG framework we presented here has quantitatively suggested that the beginning of a new wave emerges from new dominant mutations in the spike protein. However, the microscopic mechanisms behind the spreading of dominant mutations, as well as behind the end of waves, are still unknown. These aspects can be further studied following our approach, helping us to better understand the dynamics of epidemiological waves. Furthermore, the eRG can be extended to include geographical diffusion and the effect of human behavior, which are known to strongly affect the diffusion of the infections. Besides advancements in our knowledge of epidemic diffusion, which is of paramount importance for the preparedness to future pandemics, the machine learning approach we present can be used as an early warning system for the emergence of new waves. This system, which is currently under development, will be able to detect as soon as possible the emergence of a new variant of concern.

All in all, we presented a novel and self-consistent description of the diffusion of pandemics, characterized by the economy of the free parameters and by the insight on the origin of multi-wave dynamics. This is a crucial step toward the control and prevention of future pandemics caused by infectious agents.

Acknowledgments We thank the members of the team that allowed us to obtain all the novel results summarized in this chapter. They comprise theoretical physicists F. Sannino, S. Hohenegger, M. della Morte, and C. Cot; experimental particle physicists F. Conventi, F. Cirotto, and A. Giannini; computer and data scientists M. Óskarsdóttir, A.S. Islind, and A. de Hoffer; and biologists M.L. Chiusano and A. Cimarelli.

References

1. Bailey, N.: *The Mathematical Theory of Infectious Diseases*, 2nd edn. Hafner, New York (1975)
2. Balabdaoui, F., Mohr, D.: Age-stratified discrete compartment model of the COVID-19 epidemic with application to Switzerland. *Sci. Rep.* **10**, 21306 (2020)
3. Cacciapaglia, G., Cot, C., de Hoffer, A., Hohenegger, S., Sannino, F., Vatani, S.: Epidemiological theory of virus variants. *Physica A Stat. Mech. Appl.* **596**, 127071 (2022). <https://doi.org/10.1016/j.physa.2022.127071>
4. Cacciapaglia, G., Cot, C., Della Morte, M., Hohenegger, S., Sannino, F., Vatani, S.: The field theoretical ABC of epidemic dynamics (2021)
5. Cacciapaglia, G., Cot, C., Islind, A.S., Óskarsdóttir, M., Sannino, F.: Impact of us vaccination strategy on COVID-19 wave dynamics. *Sci. Rep.* **11**(1), 1–11 (2021)
6. Cacciapaglia, G., Cot, C., Islind, A.S., Óskarsdóttir, M., Sannino, F.: Impact of us vaccination strategy on COVID-19 wave dynamics. *Sci. Rep.* **11**, 10960 (2021). <https://doi.org/10.1038/s41598-021-90539-2>
7. Cacciapaglia, G., Cot, C., Sannino, F.: Mining google and apple mobility data: Temporal anatomy for COVID-19 social distancing. *Sci. Rep.* **11**, 4150 (2020). <https://doi.org/10.1038/s41598-021-83441-4>

8. Cacciapaglia, G., Cot, C., Sannino, F.: Second wave COVID-19 pandemics in europe: A temporal playbook. *Sci. Rep.* **10**, 15514 (2020). <https://doi.org/10.1038/s41598-020-72611-5>
9. Cacciapaglia, G., Cot, C., Sannino, F.: Multiwave pandemic dynamics explained: How to tame the next wave of infectious diseases. *Sci. Rep.* **11**, 6638 (2021). <https://doi.org/10.1038/s41598-021-85875-2>
10. Cacciapaglia, G., Hohenegger, S., Sannino, F.: Effective mathematical modelling of health passes during a pandemic. *Sci. Rep.* **12**, 6989 (2022)
11. Cacciapaglia, G., Sannino, F.: Interplay of social distancing and border restrictions for pandemics (COVID-19) via the epidemic Renormalisation Group framework. *Sci. Rep.* **10**, 15828 (2020). <https://doi.org/10.1038/s41598-020-72175-4>
12. Cacciapaglia, G., Sannino, F.: Evidence for complex fixed points in pandemic data. *Front. Appl. Math. Stat.* **7**, 659580 (2021). <https://doi.org/10.3389/fams.2021.659580>
13. Cardy, J.L., Grassberger, P.: Epidemic models and percolation. *J. Phys. A Math. General* **18**(6), L267–L271 (1985). <https://doi.org/10.1088/0305-4470/18/6/001>
14. Cohen, A.G., Georgi, H.: Walking Beyond the Rainbow. *Nucl. Phys. B* **314**, 7–24 (1989). [https://doi.org/10.1016/0550-3213\(89\)90109-0](https://doi.org/10.1016/0550-3213(89)90109-0)
15. de Hoffer, A., Vatani, S., Cot, C., et al.: Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for COVID-19. *Sci. Rep.* **12**, 9275 (2022)
16. Della Morte, M., Orlando, D., Sannino, F.: Renormalization Group Approach to Pandemics: The COVID-19 Case. *Front. Phys.* **8**, 144 (2020). <https://doi.org/10.3389/fphy.2020.00144>
17. Della Morte, M., Sannino, F.: Renormalization group approach to pandemics as a time-dependent sir model. *Front. Phys.* **8** (2021). <https://doi.org/10.3389/fphy.2020.591876>
18. Doi, M.: Second quantization representation for classical many-particle system. *J. Phys. A Math. Gen.* **9**, 1465 (1976). <https://iopscience.iop.org/article/10.1088/0305-4470/9/9/008>
19. Doi, M.: Stochastic theory of diffusion-controlled reaction. *J. Phys. A Math. Gen.* **9**, 1479 (1976). [https://doi.org/10.1016/S0378-4371\(03\)00458-8](https://doi.org/10.1016/S0378-4371(03)00458-8)
20. Domb, C.: Fluctuation phenomena and stochastic processes. *Nature* **184**, 509–12 (1959). <https://doi.org/10.1038/184509a0>
21. Elbe, S., Buckland-Merret, G.: Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global Chall.* **1**, 33–46 (2017). <https://doi.org/10.1002/gch2.1018>
22. Essam, J.W.: Percolation theory. *Rep. Prog. Phys.* **43**, 833 (1980). <https://iopscience.iop.org/article/10.1088/0034-4885/43/7/001/pdf>
23. Ghostine, R., Gharamti, M.E., Hassouny, S., Hoteit, I.: An extended seir model with vaccination for forecasting the COVID-19 pandemic in saudi arabia using an ensemble Kalman filter. *Mathematics* **9**, 636 (2021). <https://doi.org/10.3390/math9060636>
24. Giordano, G., Colaneri, M., Di Filippo, A., Blanchini, F., Bolzen, F., De Nicolao, G., Sacchi, P., Colaneri, P., Bruno, R.: Modeling vaccination rollouts, sars-cov-2 variants and the requirement for non-pharmaceutical interventions in italy. *Nat. Med.* (2021). <https://doi.org/10.1038/s41591-021-01334-5>
25. Grassberger, P.: On the critical behavior of the general epidemic process and dynamical percolation. *Math. Biosci.* **63**(2), 157–172 (1983). [https://doi.org/10.1016/0025-5564\(82\)90036-0](https://doi.org/10.1016/0025-5564(82)90036-0) <http://www.sciencedirect.com/science/article/pii/0025556482900360>
26. Hamer, W.: Age-incidence in relation with cycles of disease prevalence. *Trans. Epidemiol. Soc. Lond.* **15**, 64–77 (1896)
27. Hamer, W.: Epidemic disease in England: The evidence of variability and of persistency of type; Lecture 1. *Lancet*, 569–574 (1906)
28. Hamer, W.: Epidemic disease in England: The evidence of variability and of persistency of type; Lecture 2. *Lancet*, 655–662 (1906)
29. Hamer, W.: Epidemic disease in England: The evidence of variability and of persistency of type; Lecture 3. *Lancet*, 733–739 (1906)
30. Hethcote, H.W.: The mathematics of infectious diseases. *SIAM Rev.* **42**(4) (2000). <https://doi.org/10.1137/S0036144400371907>
31. Holdom, B.: Raising condensates beyond the ladder. *Phys. Lett. B* **213**, 365–369 (1988). [https://doi.org/10.1016/0370-2693\(88\)91776-5](https://doi.org/10.1016/0370-2693(88)91776-5)

32. Holdom, B.: Continuum limit of quenched theories. *Phys. Rev. Lett.* **62**, 997 (1989). <https://doi.org/10.1103/PhysRevLett.62.997>
33. Kermack, W.O., McKendrick, A., Walker, G.T.: A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A* **115**, 700–721 (1927). <https://doi.org/10.1098/rspa.1927.0118>
34. Konings, F., Perkins, M.D., Kuhn, J.H., Pallen, M.J., Alm, E.J., Archer, B.N., Barakat, A., Bedford, T., Bhiman, J.N., Caly, L., Carter, L.L., Cullinane, A., de Oliveira, T., Druce, J., Masry, I.E., Evans, R., Gao, G.F., Gorbaleyna, A.E., Hamblion, E., Herring, B.L., Hodcroft, E., Holmes, E.C., Kakkar, M., Khare, S., Koopmans, M.P.G., Korber, B., Leite, J., MacCannell, D., Marklewitz, M., Maurer-Stroh, S., Rico, J.A.M., Munster, V.J., Neher, R., Munnink, B.O., Pavlin, B.I., Peiris, M., Poon, L., Pybus, O., Rambaut, A., Resende, P., Subissi, L., Thiel, V., Tong, S., van der Werf, S., von Gottberg, A., Ziebuhr, J., Kerkhove, M.D.V.: Sars-cov-2 variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiol.* **6**, 821–823 (2021). <https://doi.org/10.1038/s41564-021-00932-w>
35. Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., Hastie, K.M., Parker, M.D., Partridge, D.G., Evans, C.M., Freeman, T.M., de Silva, T.I., Angyal, A., Brown, R.L., Carrilero, L., Green, L.R., Groves, D.C., Johnson, K.J., Keeley, A.J., Lindsey, B.B., Parsons, P.J., Raza, M., Rowland-Jones, S., Smith, N., Tucker, R.M., Wang, D., Wyles, M.D., McDanal, C., Perez, L.G., Tang, H., Moon-Walker, A., Whelan, S.P., LaBranche, C.C., Saphire, E.O., Montefiori, D.C.: Tracking changes in sars-cov-2 spike: Evidence that d614g increases infectivity of the COVID-19 virus. *Cell* **182**(4), 812–827.e19 (2020). <https://doi.org/10.1016/j.cell.2020.06.043>
36. Latif, A.A., Mullen, J.L., Alkuzweny, M., Tsueng, G., Cano, M., Haag, E., Zhou, J., Zeller, M., Hufbauer, E., Matteson, N., Wu, C., Andersen, K.G., Su, A.I., Gangavarapu, K., Hughes, L.D., the Center for Viral Systems Biology: AY4.2 Lineage Report. outbreak.info (2021). <https://outbreak.info/situation-reports?pango=AY4.2>
37. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk* **163**, 845–848 (1965)
38. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Cybern. Control Theory* **10**, 707–710 (1966)
39. Mahase, E.: Covid-19: What have we learnt about the new variant in the uk? *BMJ* **371** (2020). <https://doi.org/10.1136/bmj.m4944>
40. McKendrick, A.: The rise and fall of epidemics. *Paludism (Trans. Committee Study Malaria India)* **1**, 54–66 (1912)
41. McKendrick, A.: Studies on the theory of continuous probabilities, with special reference to its bearing on natural phenomena of a progressive nature. *Proc. Lond. Math. Soc.* **13**, 401–416 (1914)
42. McKendrick, A.: Applications of mathematics to medical problems. *Proc. Edinb. Math. Soc.* **44**, 98–130 (1926)
43. Meng, X., Cai, Z., Dui, H., Cao, H.: Vaccination strategy analysis with SIRV epidemic model based on scale-free networks with tunable clustering. *IOP Confer. Ser. Mater. Sci. Eng.* **1043**(3), 032012 (2021). <https://doi.org/10.1088/1757-899x/1043/3/032012>
44. Pater, A.A., Bosmeny, M.S., Barkau, C.L., Ovington, K.N., Chilamkurthy, R., Parasrampuria, M., Eddington, S.B., Yinusa, A.O., White, A.A., Metz, P.E., Sylvain, R.J., Hebert, M.M., Benzingier, S.W., Sinha, K., Gagnon, K.T.: Emergence and evolution of a prevalent new sars-cov-2 variant in the united states. *bioRxiv* (2021). <https://doi.org/10.1101/2021.01.11.426287>
45. Peliti, L.: Path integral approach to birth-death processes on a lattice. *J. Phys. France (Paris)* **46**, 1469–1483 (1985). <https://doi.org/10.1051/jphys:019850046090146900>
46. Perc, M., Jordan, J.J., Rand, D.G., Wang, Z., Boccaletti, S., Szolnoki, A.: Statistical physics of human cooperation. *Phys. Rep.* **687**, 1–51 (2017). <https://doi.org/10.1016/j.physrep.2017.05.004>
47. Plante, J.A., Liu, Y., Liu, J., Xia, H., Johnson, B.A., Lokugamage, K.G., Zhang, X., Muruato, A.E., Zou, J., Fontes-Garfias, C.R., Mirchandari, D., Scharton, D., Bilello, J.P., Ku, Z., An, Z., Kalveram, B., Freiberg, A.N., Menachery, V.D., Xie, X., Plante, K.S., Weaver, S.C., Shi, P.Y.: Spike mutation d614g alters sars-cov-2 fitness. *Nature* **592**, 116–121 (2021). <https://doi.org/10.1038/s41586-020-2895-3>

48. Plumptre, A.J., Baisero, D., Belote, R.T., Vázquez-Domínguez, E., Faurby, S., Jędrzejewski, W., Kiara, H., Kühl, H., Benítez-López, A., Luna-Aranguré, C., Voigt, M., Wich, S., Wint, W., Gallego-Zamorano, J., Boyd, C.: Where might we find ecologically intact communities? *Front. Forests Glob. Change* **4** (2021). <https://doi.org/10.3389/ffgc.2021.626635>. <https://www.frontiersin.org/articles/10.3389/ffgc.2021.626635>
49. Priesemann, V., Brinkmann, M.M., Ciesek, S., Cuschieri, S., Czypionka, T., Giordano, G., Gurdasani, D., Hanson, C., Hens, N., Iftekhar, E., Kelly-Irving, M., Klimek, P., Kretzschmar, M., Peichl, A., Perc, M., Sannino, F., Schernhammer, E., Schmidt, A., Staines, A., Szczurek, E.: Calling for pan-European commitment for rapid and sustained reduction in SARS-COV-2 infections. *Lancet* **397**(10269), 92–93 (2021)
50. Rambaud, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D.L., Volz, E.: Preliminary genomic characterisation of an emergent sars-cov-2 lineage in the uk defined by a novel set of spike mutations. COVID-19 Genomics Consortium UK (CoG-UK) Report (2020). <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-SARS-COV-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563>
51. Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G.: A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020). <https://doi.org/10.1038/s41564-020-0770-5>
52. Rasigade, J.P., Barray, A., Shapiro, J.T., Coquisart, C., Vigouroux, Y., Bal, A., Destras, G., Vanhems, P., Lina, B., Josset, L., Wirth, T.: A viral perspective on worldwide non-pharmaceutical interventions against COVID-19. *medRxiv* (2020). <https://doi.org/10.1101/2020.08.24.20180927>
53. Reed, K.: Viral Zoonoses. Reference Module in Biomedical Sciences, pp. B978–0–12–801238–3.95729–5 (2018)
54. Ross, R.: The Prevention of Malaria, 2nd edn. John Murray, London (1911)
55. Ross, R.: An application of the theory of probabilities to the study of *a priori* pathometry: Part I. *Proc. Roy. Soc. Lond. A* **92**, 204–230 (1916)
56. Ross, R., Hudson, H.: An application of the theory of probabilities to the study of *a priori* pathometry: Part II. *Proc. Roy. Soc. Lond. A* **93**, 212–225 (1916)
57. Ross, R., Hudson, H.: An application of the theory of probabilities to the study of *a priori* pathometry: Part III. *Proc. Roy. Soc. Lond. A* **93**, 225–240 (1916)
58. Sabino, E.C., Buss, L.F., Carvalho, M.P., Prete Jr, C.A., Crispim, M.A., Frajji, N.A., Pereira, R.H., Parag, K.V., da Silva Peixoto, P., Kraemer, M.U., Oikawa, M.K., Salomon, T., Cucunuba, Z.M., Castro, M.C., Aruska de Souza Santos, A., Nascimento, V.H., Pereira, H.S., Ferguson, N.M., Pybus, O.G., Kucharski, A., Busch, M.P., Dye, C., Faria, N.R.: Resurgence of COVID-19 in manaus, brazil, despite high seroprevalence. *Lancet* **397**, 452–455 (2021). [https://doi.org/10.1016/S0140-6736\(21\)00183-5](https://doi.org/10.1016/S0140-6736(21)00183-5)
59. Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R.: Viral mutation rates. *J. Virol.* **84**(19), 9733–9748 (2010). <https://doi.org/10.1128/JVI.00694-10>
60. Shu, Y., McCauley, J.: Gisaid: Global initiative on sharing all influenza data – from vision to reality. *EuroSurveillance* **22** (13) (2017). <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494>
61. Stauffer, D.: Scaling theory of percolation clusters. *Phys. Rep.* **54**, 1–74 (1979)
62. Taubenberger, J.K., Morens, D.M.: 1918 influenza: The mother of all pandemics. *Rev. Biomed.* **17**(1), 69–79 (2006)
63. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A.J., Engelbrecht, S., Van Zyl, G., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, M., Korsman, S., Davies, M.A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C.K., Sewell, B.T., Lourenço, J., Alcantara, L.C.J., Pond, S.L.K., Weaver, S., Martin, D., Lessells, R.J., Bhiman, J.N., Williamson, C., de Oliveira, T.: Emergence

- and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (sars-cov-2) lineage with multiple spike mutations in South Africa. medRxiv (2020). <https://doi.org/10.1101/2020.12.21.20248640>
64. Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C.V., Boyd, O., Loman, N.J., McCrone, J.T., Gonçalves, S., Jorgensen, D., Myers, R., Hill, V., Jackson, D.K., Gaythorpe, K., Groves, N., Sillitoe, J., Kwiatkowski, D.P., Flaxman, S., Ratmann, O., Bhatt, S., Hopkins, S., Gandy, A., Rambaut, A., Ferguson, N.M.: Transmission of sars-cov-2 lineage b.1.1.7 in England: Insights from linking epidemiological and genetic data. medRxiv (2021). <https://doi.org/10.1101/2020.12.30.20249034>
65. Wang, Z., Andrews, M.A., Wu, Z.X., Wang, L., Bauch, C.T.: Coupled disease–behavior dynamics on complex networks: A review. *Phys. Life Rev.* **15**, 1 – 29 (2015). <https://doi.org/10.1016/j.plrev.2015.07.006>
66. Wang, Z., Bauch, C.T., Bhattacharyya, S., d'Onofrio, A., Manfredi, P., Perc, M., Perra, N., Salathé, M., Zhao, D.: Statistical physics of vaccination. *Phys. Rep.* **664**, 1 – 113 (2016). <https://doi.org/10.1016/j.physrep.2016.10.006>. <http://www.sciencedirect.com/science/article/pii/S0370157316303349>
67. Who “tracking sars-cov-2 variants” page. <https://www.who.int/en/activities/tracking-SARS-COV-2-variants/>
68. Wilson, K.G.: Renormalization group and critical phenomena. 1. Renormalization group and the Kadanoff scaling picture. *Phys. Rev. B* **4**, 3174–3183 (1971). <https://doi.org/10.1103/PhysRevB.4.3174>
69. Wilson, K.G.: Renormalization group and critical phenomena. 2. Phase space cell analysis of critical behavior. *Phys. Rev. B* **4**, 3184–3205 (1971). <https://doi.org/10.1103/PhysRevB.4.3184>
70. Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G., Jiang, T.: Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**(3), 325–328 (2020). <https://doi.org/10.1016/j.chom.2020.02.001>

Chapter 4

Contact Matrices in Compartmental Disease Transmission Models



Jesse Knight and Sharmistha Mishra

4.1 Introduction

At the core of mathematical models of infectious disease transmission is the “force of infection” (incidence per susceptible) equation, which defines the average rate at which susceptible individuals become infected:

$$\lambda(t) = C\beta \frac{I(t)}{N} \quad (4.1)$$

where $I(t)$ is the number of infectious individuals, N is the total population, β is the average probability of transmission per contact, and C is the average rate of contact between individuals per unit time.¹ In (4.1), the population is assumed to be homogeneous, i.e. having equal susceptibility to infection, equal infectiousness once infected, and equal contact rates. In reality, populations are rarely homogeneous, and differential risk (acquisition and/or transmission) can strongly influence epidemic dynamics [10]. So, we are often interested in modelling transmission within *stratified* populations. Such stratifications can be included as additional compartments within compartmental models.

¹ Equation (4.1) assumes “frequency-dependent” transmission, which does not scale with population density; an alternate assumption is called “density-dependent” transmission, which replaces the population size N with an area A in (4.1), and changes the interpretation of C [5]. We assume frequency-dependent transmission throughout this chapter, because the survey-derived contact data used to inform C (see Sect. 4.2) are taken as fixed, as opposed to scaling with population density N/A .

J. Knight (✉) · S. Mishra

MAP Centre for Urban Health Solutions, Unity Health Toronto, Toronto, ON, Canada

e-mail: jesse.knight@mail.utoronto.ca; sharmistha.mishra@utoronto.ca

For a stratified population, the force of infection for group (stratum) i is defined via the sum of contributions from all groups i' , including from other individuals in the same group $i' = i$:

$$\lambda_i(t) = \sum_{i'} C_{ii'} \beta_{ii'} \frac{I_{i'}(t)}{N_{i'}} \quad (4.2)$$

where $C_{ii'}$ is the average rate that individuals in group i contact individuals in group i' and $\beta_{ii'}$ is the average probability of transmission per contact from individuals in group i' to individuals in group i . Thus, C and β are now matrices. Although there are many interesting factors that can contribute to differences across strata or “heterogeneity” in β , the focus of this chapter will be on contact matrices C and applications in the context of compartmental transmission dynamics models. Specifically, we will focus on data and methods to define contact matrices for populations stratified by age and geographic patches.

The remainder of this chapter is organized as follows: First, in Sect. 4.2, we introduce a motivating example for contact matrices to help illustrate the results of methodology throughout the chapter. Then, in Sect. 4.3, we summarize various data and assumptions which can be used to define contact matrices. Next, in Sect. 4.4, we review the basic properties and representations of contact matrices. After, in Sect. 4.5, we explore the problem of restratifying contact matrices. Finally, in Sect. 4.6, we explore a framework for integrating mobility data into contact matrices. For reference, Table 4.1 summarizes our notation for key indices and variables. Note that matrix indices ij denote row i and column j , for consistency with R code.

Table 4.1 Notation

	Symbol	Definition
Indices	a	Self age group
	a'	Other age group
	g	Self patch
	g'	Other patch
	y	Contact type
	t	Time
Variables	N	Population size
	ϕ	Probability of contact
	C	Contacts per person
	Γ	Intrinsic connectivity matrix
	X	Total contacts in population
	M	Mobility matrix

4.2 Motivating Example

To illustrate several methods and their results throughout the chapter, we draw on results from our previous work [16]. This work sought to define contact matrices for the population of Ontario, Canada (population 15 million), for the purposes of SARS-CoV-2 transmission modelling. The contact matrices were stratified by four factors:

- Ten quasi-geographic **patches**, reflecting collections (deciles) of neighbourhoods (forward sortation areas) ranked by cumulative COVID-19 incidence between 15 January 2022 and 28 March 2021 (Fig. 4.1)
- Irregular **age** groups, reflecting past COVID-19 vaccination eligibility in Ontario: 0–11, 12–15, 16–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80+
- Contact **type** (location), matching the four locations used in [23]: home, work, school, and others
- **Time**, reflecting monthly changes in mobility patterns during 2020

To define these matrices, we incorporated age-stratified contact matrices from [23] (Fig. 4.2) and built upon patch-mixing methods from [2], as described later in the chapter.

4.3 Defining Contact Matrices

Just as compartmental models combine necessary simplifying assumptions with key sources of epidemiological data, contact matrices are defined using some combination of data and assumptions. In this section, we summarize key sources of contact data and common assumptions, starting with the question . . .

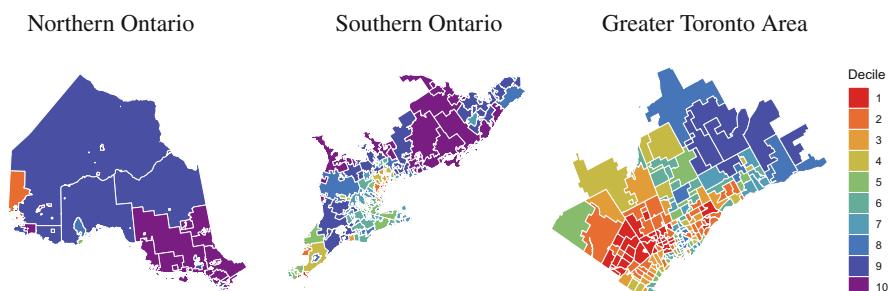


Fig. 4.1 Ontario neighbourhoods ($N = 513$), stratified by decile rank in cumulative COVID-19 cases between 15 January 2020 and 28 March 2021. Decile rank was used to group neighbourhoods into ten patches for transmission modelling

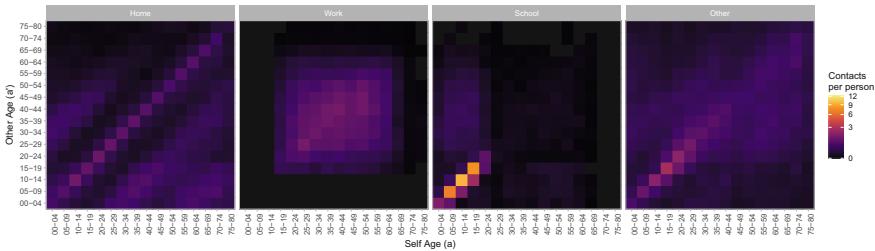


Fig. 4.2 Age-stratified contact matrices $C_{ii'y}$ from [23], further stratified by contact type defined by contact location Contact matrices for Canada (per-person scale). Colour scales are square-root transformed to improve perception of smaller values

4.3.1 What Is a Contact?

For the purposes of transmission modelling, a “contact” refers to an exposure that could potentially lead to a transmission event. Therefore, a contact is defined based on the pathogen(s) and mode(s) of transmission captured in the modelled system. For example, definitions of a contact could include (but are not limited to):

- **Proximity:** an extended time in the same general vicinity, where transmission can occur via aerosols, or a shorter time at a closer distance, where transmission can occur via larger droplets,² e.g. respiratory pathogens like tuberculosis and SARS-CoV-2
- **Direct contact:** skin-to-skin contact, or, in some cases, touching shared objects and environmental surfaces called “fomites”, e.g. certain bacteria like methicillin-resistant *Staphylococcus aureus* (MRSA)
- **Sexual contact:** direct contact during sexual activity at the site of mucous membranes (genitals, rectum, oral) and/or via exposure of specific mucosal tissues with genital fluid, e.g. sexually transmitted infections like syphilis

Many pathogens can have more than one mode of transmission, and the probability of transmission per contact β can vary both between and within modes of transmission. For example, for the same pair of individuals (one susceptible, one infectious), singing indoors without masks for multiple hours poses a higher risk of SARS-CoV-2 transmission, than if both individuals were walking in an outdoor market for a few minutes; similarly, in the absence of condoms and HIV pre-exposure prophylaxis, anal receptive sex poses a higher risk of HIV acquisition than anal insertive sex. Thus, different types of contacts are sometimes explicitly modelled within the same system.

² The distinction between aerosols and droplets falls along a spectrum but is mainly defined by the ability of aerosols to, unlike droplets, remain suspended in the air for several seconds or more.

4.3.2 Sources of Contact Data

Empirical survey data on different types of contacts are the key data source for contact matrices. These surveys are largely administered as weekly or daily diaries. The data are self-reported answers to questions about the numbers of contacts in a given time period (usually the previous day), stratified by type and/or location of contacts (e.g. household, workplace, school, etc.), and the attributes of the persons with whom each contact occurred (e.g. age, gender, vaccination status, etc.). These attributes are critical to developing contact *matrices* because they explicitly characterize “who contacts whom” with respect to the modelled population stratifications.

For data on non-sexual contacts, the most well-known source of survey-based data is POLYMOD (Improving Public Health Policy in Europe through Modelling and Economic Evaluation of Interventions for the Control of Infectious Diseases) [19]. The POLYMOD study was explicitly designed to inform transmission modelling and used self-reported survey data to estimate epidemiologically relevant contact patterns from representative samples of 8 different European countries, totalling 7290 respondents and 97,904 reported contacts. Within POLYMOD, contacts were defined as [19]:

either skin-to-skin contact such as a kiss or handshake (a physical contact), or a two-way conversation with three or more words in the physical presence of another person but no skin-to-skin contact (a nonphysical contact).

Such contacts are thus relevant to transmission of several pathogens, although the per-contact transmission risk β may vary from pathogen to pathogen.

There are also a large number of surveys outside of Europe that have collected data on non-sexual contacts, such as in South Africa [14, 15]. In Canada, the most common source of contact data for modelling respiratory pathogens is the CONNECT study [9]. During the COVID-19 pandemic, additional regional surveys were also developed to inform the study of local transmission dynamics of SARS-CoV-2, such as BCMIX in British Columbia [1] and a repeat cross-sectional survey in Ontario [8].

Surveys on sexual contacts ask about self-reported numbers and types of sexual partnerships, as well as numbers and types of sex act within those partnership; some also ask about partnership attributes to help generate partnership matrices. Sexual behaviour surveys are routinely administered in many countries as part of demographic health and/or related national surveys, although self-reported data on sexual activity is often subject to measurement biases [6]. Many such surveys are conducted among populations experiencing disproportionate risk of sexually transmitted infections, especially HIV [4, 7]. Such surveys require community engagement, community-driven efforts, and complex survey sampling methodologies to reach populations at disproportionate risk, who simultaneously face marginalization and stigma (such as individuals engaged in sex work and men who have sex with men) [4, 7].

4.3.3 Assumptions and Parametric Forms

In some cases, context-specific contact data are not available, or do not contain sufficient information on the attributes of contacts, such that C_i is known, but $C_{i'}$ is not. Sometimes even C_i must be assumed. In these cases, we must make additional assumptions about “who contacts whom”. Such assumptions are usually specified using one of the following parametric forms.

The simplest assumption is “random mixing”, also called “proportional mixing”. In random mixing, the probability $\phi_{ii'}$ that an individual in stratum i forms a *given* contact with an individual from stratum i' is proportional to the total number of contacts “offered” by individuals in stratum i' :

$$\phi_{ii'} = \frac{C_{i'} N_{i'}}{\sum_{i'} C_{i'} N_{i'}} \quad (4.3)$$

Thus, $\phi_{ii'}$ does not depend on i and so each row is identical. At the population level (see Sect. 4.4), proportional mixing is defined by the outer product of $N_i C_i$ with itself, normalized by the total number of contacts:

$$X_{ii'} = \frac{C_i N_i \cdot C_{i'} N_{i'}}{\sum_{i'} C_{i'} N_{i'}} \quad (4.4)$$

Along the spectrum of mixing is the assumption of “assortative mixing”, wherein each individual is more likely to contact other individuals in the same stratum. At the extreme, $\phi_{ii'}$ can be an identity matrix:

$$\phi_{ii'} = \delta_{ii'}, \quad \delta_{ii'} = \begin{cases} 1 & i = i' \\ 0 & i \neq i' \end{cases} \quad (4.5)$$

such that individuals *only* contact other individuals in the same stratum. For less extreme mixing, the following equation can be used [20]:

$$\phi_{ii'} = (\epsilon) \delta_{ii'} + (1 - \epsilon) \frac{C_{i'} N_{i'}}{\sum_{i'} C_{i'} N_{i'}} \quad (4.6)$$

which “interpolates” between the two extremes of complete assortativity ($\epsilon = 1$) and random mixing ($\epsilon = 0$). This popular approach (4.6) is guaranteed to satisfy mixing matrix constraints (see Sect. 4.4), but it cannot represent complex mixing patterns.

For more complex mixing patterns, a highly general approach was introduced in [18]. This approach defines contact patterns at the population level $X_{ii'}$, starting from random mixing (4.4). Then, a symmetric matrix $\theta_{ii'}$ is used to specify the disproportionate log-odds of contacts between strata i and i' :

$$X_{ii'} = \frac{C_i N_i \cdot C_{i'} N_{i'}}{\sum_{i'} C_{i'} N_{i'}} \exp(\theta_{ii'}) \quad (4.7)$$

A framework for estimating elements of $\theta_{ii'}$ from empirical data is also developed in [18]. Notably, (4.7) changes the numbers of contacts per person C_i reflected in $X_{ii'}$, and no closed-form definition of $X_{ii'}$ can maintain fixed C_i for arbitrary $\theta_{ii'}$. However, an iterative proportional fitting procedure [24] can resolve $X_{ii'}$ which maintains the original C_i as follows:

$$X_{ii'}^{(n+1)} = X_{ii'}^{(n)} \frac{C_k N_k}{\sum_k X_{ii'}^{(n)}} \quad k = \begin{cases} i & n \text{ is even} \\ i' & n \text{ is odd} \end{cases} \quad (4.8)$$

which typically converges within $n = 20$ iterations. Other parametric approaches are also possible, typically guided by specific knowledge and assumptions about the modelled contacts and population strata.

Finally, it is worth reiterating the fundamental assumption of compartmental models: that each population stratification (“compartment”) is itself homogeneous. Thus, the average contact rates captured in contact matrices are assumed to apply equally to all individuals within the relevant strata.

4.3.4 Example

Returning to our motivating example (Sect. 4.2), the main source of contact data we use was originally obtained from the POLYMOD study [19]. Despite the great design of this study, it was not clear how POLYMOD data might generalize to other populations, having unique demographics, household structures, education, and labour characteristics. This generalization problem was tackled in [22] and later updated in [23]. These works incorporate the POLYMOD data alongside several sources of more widely available demographic data in a Bayesian hierarchical model, allowing estimation of contact matrices for 169 countries not captured in POLYMOD. These estimates for all 177 countries are happily available online, stratified by 5-year age groups.³

As the starting point for developing our age-/patch-/type-stratified contact matrices, we extracted the matrices for Canada from this database, illustrated in Fig. 4.2. We observe several general trends in these matrices. First, across all contact types, we see a prominent diagonal, reflecting increased likelihood of contacting others of the same age. At home, the diagonal likely reflects spousal partnerships and siblings of similar age, while at school, the diagonal likely reflects cohorts of the same grade. The home contacts also feature multiple off-diagonal “ridges”, likely

³ github.com/kieshaprem/synthetic-contact-matrices.

reflecting intergenerational contacts. Work contacts are mainly concentrated among working aged adults, and likewise for school contacts, with the exception of a small number of working aged adults, likely reflecting school faculty and staff.

4.4 Properties of Contact Matrices

Contact matrices can be defined and studied on three major scales:

Contact Matrix Scales

1. **Probability:** $\phi_{ii'}$ represents the probability that individuals in group i will form a *given* contact with an individual in group i' .
2. **Per-Person:** $C_{ii'}$ represents the *average number* of contacts formed by individuals in group i with individuals in group i' . We therefore have $C_{ii'} = C_i \phi_{ii'}$, where C_i is the average number of contacts formed by individuals in group i overall.
3. **Population:** $X_{ii'}$ represents the total number of contacts formed between all individuals in group i with all individuals in group i' . We therefore have $X_{ii'} = N_i C_{ii'}$, where N_i is the population size (or proportion) of group i .

The relationships between these three scales can be summarized as:

$$X_{ii'} = N_i C_{ii'} = N_i C_i \phi_{ii'} \quad (4.9)$$

Thus, it is relatively straightforward and often useful to switch between these scales for different tasks. Contact matrices also have three important constraints. These constraints should be verified when defining matrices (see Sect. 4.3), and especially when contact matrices change over time, in response to changes in N_i , C_i , and/or $\phi_{ii'}$. For example, the population proportions within each stratum N_i may change over time due to demographic shifts or differential infection-attributable mortality; contact rates C_i may change over time in response to changing risk perception or public health recommendations such as physical distancing; and mixing patterns $\phi_{ii'}$ may change over time due to changing social factors or, again, in response to public health interventions.

Contact Matrix Constraints

Here the constraints are defined on the probability scale, but equivalent constraints can be defined on all three scales.

- Element Bounds:** Each matrix element must be a valid probability:

$$\phi_{ii'} \in [0, 1] \quad (4.10)$$

- Row Sums:** The sum of matrix elements $\phi_{ii'}$ across each row i must be one:

$$\sum_i \phi_{ii'} = 1 \quad (4.11)$$

- Balancing:** The total number of contacts from group i to group i' must equal the total number of contacts from group i' to group i :

$$N_i C_i \phi_{ii'} = N_{i'} C_{i'} \phi_{i'i} \quad (4.12)$$

This constraint is sometimes called “reciprocity”. At the population scale, and only the population scale, this constraint means that $X_{ii'} = X_{i'i}$ —i.e. X is symmetric.

4.4.1 Balancing Contact Matrices

There are three main reasons why the balancing constraint could be unsatisfied. First, when survey data are used to define the contact matrices (see Sect. 4.3.2), contacts might not balance due to survey error, namely, sampling error (the survey did not reach a representative sample of the population), or response error (the survey respondents did not accurately recall or report their contacts). Second, as noted above, the population proportions within each stratum may change over time. Third, contact behaviour within each stratum may change over time.

Most of the parametric approaches to define contact matrices (see Sect. 4.3.3) inherently produce balanced matrices. When contact matrices are not balanced due to survey error, a popular solution is to average the imbalanced matrix $X_{ii'}^*$ with its transpose at the population scale:

$$X_{ii'} = \frac{1}{2} X_{ii'}^* + \frac{1}{2} X_{i'i}^* \quad (4.13)$$

thus ensuring the resulting $X_{ii'}$ is symmetric. This approach only works on the population scale, with one exception: when the population sizes of all strata are equal, $N_i = N_{i'}$ for all i, i' , (4.13) can also be applied on the per-person scale $C_{ii'}$.

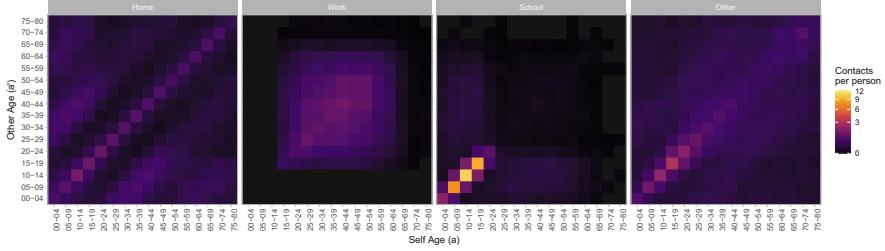


Fig. 4.3 Age-stratified intrinsic connectivity matrices $\Gamma_{ii'y}$. Contact matrices for Canada derived from [23] (per-person scale). Colour scales are square-root transformed to improve the perception of smaller values

4.4.2 Intrinsic Connectivity

The case where all strata have equal population size is related to the “intrinsic connectivity matrix” from [3], defined as:

$$\Gamma_{ii'} = C_{ii'} \frac{N}{N_{i'}} \quad (4.14)$$

This matrix $\Gamma_{ii'}$ is a useful intermediate representation of contacts on the per-person scale after removing the inherent weighting by stratum population sizes ($N_{i'}/N$).⁴ In the context of age, $\Gamma_{ii'}$ can be interpreted as the expected contact matrix for a population with a flat/rectangular demographic pyramid. Then, a contact matrix $C_{ii'}$ can easily be defined for a new population with different stratum sizes by reversing (4.14). Like $X_{ii'}$, $\Gamma_{ii'}$ should be symmetric if contacts are balanced.

4.4.3 Example

The contact matrices estimated in [23] were not guaranteed to balance, and used national-level age distributions. In our transmission modelling application, contact matrices must balance and will later be adapted to several unique age distributions (see Sect. 4.6.3). Thus, our first step in using these data is to compute a balanced intrinsic connectivity matrix $\Gamma_{ii'y}$ for each contact type y .

Applying the methods described above, namely, Eqs. (4.14) and (4.13), to the matrices from Fig. 4.2, we obtain the balanced intrinsic connectivity matrices $\Gamma_{ii'y}$ shown in Fig. 4.3. The resulting matrix is evidently symmetric, and includes slightly more contacts with the oldest age groups, but is otherwise similar to Fig. 4.2.

⁴ The intrinsic connectivity matrix Γ serves a similar role to the log-odds matrix θ from Sect. 4.3.3, but the two matrices are not the same.

4.5 Restratiyng Contact Matrices

One challenge in applying empirical contact data is that, for continuous variables that define individuals' attributes, the population stratification of the data may not match the desired stratification for analysis. For example, the contact matrices provided in [23] use regular 5-year age groups, while our motivating example requires irregular age groups, reflecting vaccine eligibility [17]. In this section, we introduce some simple techniques to restratify contact matrices to or from any stratification. We focus on age-stratified matrices, but the same techniques could be applied to matrices stratified by other variables.

4.5.1 Intuition and Equations for Restratiyng

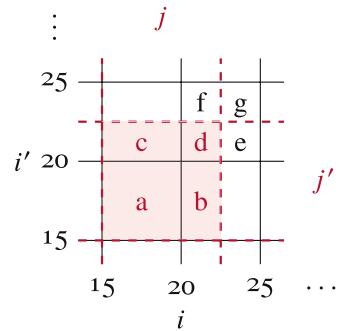
We begin by noting that restratiyng contact matrices is easier to do on the population scale $X_{ii'}$, where matrix elements $x_{ii'}$ reflect absolute numbers of contacts between groups i and i' . In this case, restratiyng from ii' to jj' involves the same operations for both $i \rightarrow j$ and $i' \rightarrow j'$, namely, summing the numbers of contacts (to form larger stratifications), or splitting them up (to form smaller stratifications). For example, restratiyng $X_{ii'}$ from 5-year age groups (ii') to 10-year age groups (jj') involves summing adjacent 2×2 blocks of $x_{ii'}$ to yield $x_{jj'}$. Thus, the magnitude of $x_{jj'}$ generally increases with larger j and with larger j' .

On the per-person and probability scales ($C_{ii'}$ and $\phi_{ii'}$), restratiyng $i' \rightarrow j'$ is the same as before, but restratiyng $i \rightarrow j$ now reflects different groups of people whose contact data are averaged—i.e. a different definition of the average person in the group. Thus, the magnitudes of $c_{jj'}$ and $\phi_{ii'}$ generally increase with larger j' , but *not* with larger j . Of course, it is always possible to convert back to per-person or probability scales after restratiyng on the population scale.

Next, we develop the equations for restratiyng. To support development of these equations, Fig. 4.4 illustrates a simple case of restratiyng from 5-year age groups to 7.5-year age groups. Intuitively, we see that the total number of contacts $x_{jj'}$ for $j = j' = [15, 22.5)$ (colour) should include from $X_{ii'}$: (a) all contacts from $i = i' = [15, 20)$, (b) some contacts from $i = [20, 25)$, $i' = [15, 20)$, (c) some contacts from $i = [15, 20)$, $i' = [20, 25)$, and (d) some contacts from $i = i' = [20, 25)$.

It is tempting to specifically use half of (b), half of (c), and one-quarter of (d), and overall this is a reasonable approach. However, this approach assumes that contact patterns are “flat” within each current stratum, such that (d) = (e) = (f) = (g), and so on. In fact, age-stratified contact matrices often include strong diagonal trends (e.g. Fig. 4.2), from which we would expect that (d), (g) > (e), (f), even though such trends may not be visible in $X_{ii'}$ due to averaging. We could also have that (d) > (g), if younger people tend to have more contacts, and so on.

Fig. 4.4 Diagram of restratification from 5-year age groups ii' to 7.5-year age groups jj'



To develop a better approach, let us consider the “true” population-level contact function $X(\mu, \mu')$, where $\mu, \mu' \in \mathbb{R}$ are *continuous* age variables. From $X(\mu, \mu')$, we can define any aggregated $X_{jj'}$ through integration:

$$X_{jj'} = \int_{\mu \in j} \int_{\mu' \in j'} X(v, v') dv dv' \quad (4.15)$$

Obtaining an exact definition of $X(\mu, \mu')$ is practically impossible. In many cases, it should be sufficient to obtain a discrete matrix $X_{uu'}$, with 1-year strata u, u' (or some equivalent base unit). Then, the integrals of (4.15) can be replaced by sums:

$$X_{jj'} = \sum_{u \in j} \sum_{u' \in j'} X_{uu'} \quad (4.16)$$

The matrix $X_{uu'}$ can then be used to reflect assumed contact patterns *within* the original strata, beyond the “flat” assumption described above.

Unfortunately, $X_{uu'}$ is difficult to infer from the source data $X_{ii'}$, because the averaging of contacts within each age group results in irrecoverable loss of information, similar to other types of signals sampled below the Nyquist frequency [21]. Figure 4.5 illustrates this challenge graphically, where we see that the age group midpoints and average contacts (X_i , dots) are not necessarily points along the true contact function (X_u , coloured curve). In particular, X_i will tend to underestimate X_u in concave down regions (peaks) and overestimate X_u in concave up regions (valleys). The underestimation of peaks is of particular concern, since groups with the highest contact rates are often key determinants of epidemic emergence and persistence [10].

To our knowledge, an unbiased estimator of $X_{uu'}$ given $X_{ii'}$ is not available, and the development of such an estimator would be a great contribution to the field. Instead, modellers typically use constant (nearest neighbour), bilinear, or bicubic interpolation, with the limitations described above.

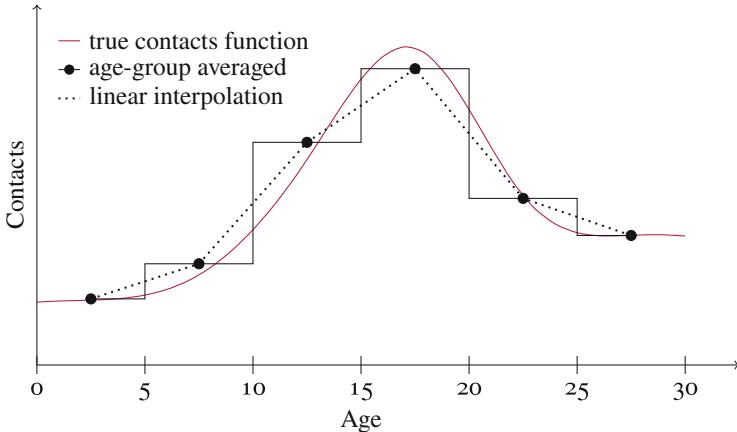


Fig. 4.5 Imperfect reconstruction of sampled contacts in one dimension

4.5.2 Example

To illustrate these techniques, we begin by restratifying the intrinsic connectivity matrices from Sect. 4.4.3 using three different interpolation techniques. Figure 4.6a compares constant, bilinear, and bicubic interpolation to estimate a 1-year matrix $\Gamma_{uu'}$ from the 5-year matrix $\Gamma_{ii'}$ (sum of four contact types, per-person scale). Figure 4.6b also shows the reconstructed input 5-year matrices $\Gamma_{jj'}$ via (4.16) with $j = i$, and Fig. 4.6c shows the reconstruction error. We can see that constant interpolation (top) allows perfect reconstruction of the input $\Gamma_{ii'}$, but qualitatively produces the least plausible 1-year matrix $\Gamma_{uu'}$, and some notable artifacts in the 2-year restratification (Fig. 4.6d). Bicubic interpolation (bottom) produces arguably the most plausible 1-year matrix $\Gamma_{uu'}$ and incurs only moderate reconstruction error, so it is likely the best choice of these three options for this task.

Recalling the irregular target age strata j defined in Sect. 4.2, we can apply the same methods as above to obtain Fig. 4.7 (per-person scale). The horizontal streaks are not an error, but rather reflect age groups of variable sizes. For example, age groups <12 and 16–39 are large, so individuals are more likely to contact individuals in these age groups, whereas age group 13–15 is much smaller, etc. Horizontal streaks appear on all three matrix scales, while symmetric vertical streaks also appear on the population scale. The matrices in Fig. 4.7 are no longer symmetric, because the new strata are different sizes. However, the matrices still reflect intrinsic connectivity for a population with flat/rectangular age demography and can still be adapted to new demography via (4.14) as described in Sect. 4.4.2.

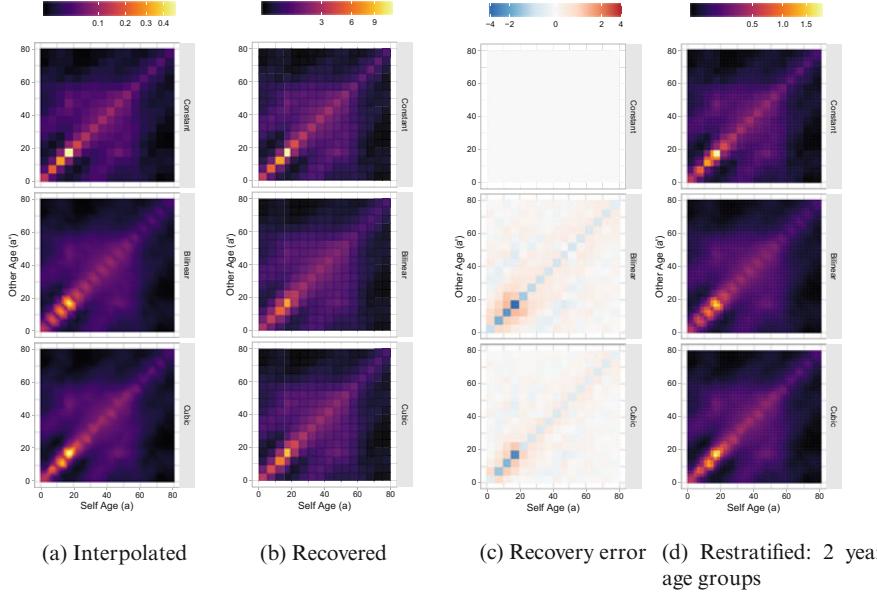


Fig. 4.6 Interpolation, recovery, and resampling of overall contact matrices using constant, bilinear, and bicubic interpolation methods Starting from contact matrices stratified by 5-year age groups (sum across contact types in Fig. 4.3), interpolation **(a)** estimates matrices stratified by 1-year age groups. This interpolation can prevent perfect recovery of the input matrix via re-summing age groups **(b)**, with errors (recovered—input) shown in **(c)**. However, some interpolation methods produce smoother and thus perhaps more plausible contact matrices when re-summing to new age groups, e.g. 2-year groups **(d)**. Contact matrices for Canada originally derived from [23] (per-person scale). Colour scales are square-root transformed to improve perception of smaller values

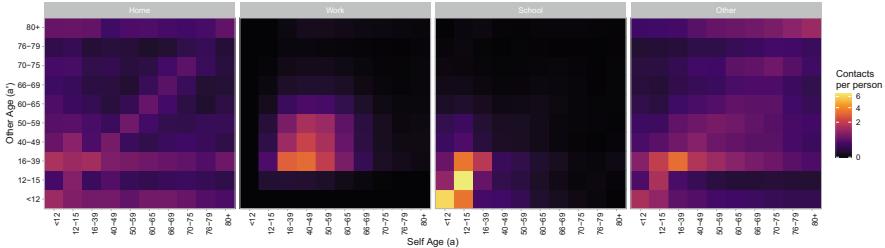


Fig. 4.7 Contact matrices resampled to the irregular age groups of interest Starting from the regular 5-year age groups in Fig. 4.3, contact matrices were restratiﬁed by transformation to the population scale; bilinear interpolation to 1-year age groups, summing to the target age groups, 0–11, 12–15, 16–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80+; and finally transformation back to the per-person scale. Contact matrices for Canada originally derived from [23]. Colour scales are square-root transformed to improve perception of smaller values

4.6 Mobility in Contact Matrices

Until now, we have mainly been discussing contact matrices stratified by age. Many models also stratify populations by so-called patches, reflecting geographic regions or community membership. Within patch-stratified models, individuals may then travel between patches permanently, such as changing residence, or, recurrently, such as for work or school. Here we focus on recurrent mobility. Such mobility can influence contact patterns, so we develop an approach to integrate mobility and contact data to inform contact matrices.

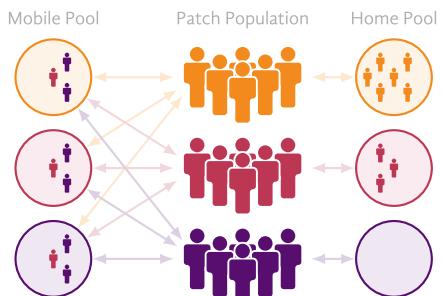
As suggested in [12], we begin by conceptualizing two mobility-related contexts (“pools”) in which contacts may be formed:

1. **Home pools:** where contacts are formed exclusively with other members of the same patch (e.g. household contacts)
2. **Mobile pools:** where contacts are formed with any individuals present in the pool (e.g. work contacts)

Figure 4.8 illustrates home and mobile pools for a simple population with three patches, where zero, half, and all individuals from patches 1 (yellow), 2 (red), and 3 (purple) are mobile, respectively. From this example, we can make a few observations. First, individuals can be mobile within their own mobile pool or travel to another mobile pool; we distinguish these cases as “mobile-at-home” versus “mobile-away”. Second, purple individuals from patch 3 can thus contact red individuals from patch 2 within mobile pool 3 (purple is *mobile-at-home*), within mobile pool 2 (red is *mobile-at-home*), or within mobile pool 1 (purple and red are both *mobile-away*). This last situation then allows potential contact formation between seemingly disconnected patches. Finally, we may wish to assume that certain types of contacts are formed only, or mostly, within one type of pool or the other (e.g. household contacts in home pools, versus work contacts in mobile pools).

In Sect. 4.6.2, we continue with the framework introduced above to develop the equations for mobility-related contact patterns. But first in Sect. 4.6.1, we introduce some sources of mobility data, and how they can be used.

Fig. 4.8 Simple example of “home” vs “mobile” pools, in a population with three patches



4.6.1 Mobility Data and Mobility Matrices

In general, different types of mobility data can be used to generate a mobility matrix $M_{gg'}$, describing the numbers or proportions of individuals in patch g who travel to patch g' per unit time. Often, these data reflect recurrent mobility, such that travelling individuals return to their home patch g each unit time. However, some data may also reflect permanent movement between patches, which can be incorporated as a change in patch membership for travelling individuals from g to g' . Also, in some cases, it may not be possible or relevant to specify the diagonal of $M_{gg'}$ (where $g = g'$), reflecting mobility within the residence patch; in other cases, this mobility data may be available and indeed important.

Common sources of anonymized mobility data are summarized in [13], such as transportation data (e.g. buses, trains, flights, etc.), mobile network data (e.g. logs of mobile phone calls and text messages), and GPS data (e.g. coordinates requested via mobile phone apps). Such data are usually sampled by convenience, as opposed to purposeful data collection from a representative sample, so non-representativeness is often a notable concern. For example, individuals accessing GPS services via navigation apps may be more likely to travel versus other individuals. Various assumptions and/or additional data can be used to try to correct for representativeness, such as computing *relative* mobility versus a reference period [16]. Other challenges include how to identify the residence patch of mobile devices in the dataset and how to define and interpret “mobility” to other patches via timestamped observations of devices in particular locations.

In the absence of context-specific mobility data, several parametric approaches have also been proposed to derive mobility matrices [25]. One popular approach is the “gravity model”, which models the number of travellers from patch g to g' as:

$$M_{gg'} = \frac{N_g^\alpha N_{g'}^{\alpha'}}{f(d_{gg'})} \quad (4.17)$$

that is, proportional to the population size in each patch N_g , $N_{g'}$ (possibly transformed via exponents α and α' , respectively) and inversely proportional to the distance between patches $d_{gg'}$ (again, possibly with transformation f). The parameters α and α' and function f can be calibrated or assumed.

4.6.2 Contact Matrices from Mobility Matrices

Returning to the task of developing contact matrices using mobility data, we begin by specifying some assumptions. First, we assume that the diagonal of the mobility matrix $M_{gg'}$ ($g = g'$) is available from the data and that these values represent the proportions of individuals from patch g who are mobile within their own mobile

pool g each day. The off-diagonal elements ($g \neq g'$) then represent the proportions of individuals from patch g who travel to mobile pool g' each day.

Second, we assume that the numbers of contacts formed by mobile individuals do not depend on the visited patch; thus, $M_{gg'}$ reflects the conditional distribution of mobile pools g' where contacts will be formed among mobile individuals from patch g . Third, we assume that mixing by residence patch within mobile pools is effectively random. Finally, we introduce a parameter $h_y \in [0, 1]$ which represents the proportion of type- y contacts that are formed in the home pool, and the remainder ($1 - h_y$) are formed with mobile pools. For example, we could have $h_y = 1$ for household contacts and $h_y = 0$ for work contacts. With these assumptions, we model the contacts formed by individuals in patch g as distributed across three situations:

Mobility-Related Situations of Contact Formation

Individuals can be modelled to form contacts in three main situations:

1. **Mobile-Away:** Individuals travelled from patch g to g' and formed contacts within mobile pool g' . Modelled proportion of contacts: $(1 - h_y)M_{gg'}, g \neq g'$.
2. **Mobile-At-Home:** Individuals formed contacts within their local mobile pool g . Modelled proportion of contacts: $(1 - h_y)M_{gg'}, g = g'$.
3. **Non-Mobile-At-Home:** Individuals formed contacts within their home pool g . Modelled proportion of contacts: $h_y\delta_{gg'}$.

Now we develop the equations for mobility-based contact matrices. We ignore age stratification for now but will add it below in Sect. 4.6.3. Consider the mobile pool in a given patch g^* . The total number of individuals from patch g who are present in mobile pool g^* can be defined using the mobility matrix $M_{gg'}$ as:

$$N_g^* = N_g M_{gg^*} \quad (4.18)$$

Within this pool g^* , we assume that mixing by patch is random, so the total numbers of type- y contacts between members of different patches can be defined as:

$$X_{gg'y}^* = (1 - h_y) \frac{C_{gy} N_g^* C_{g'y} N_{g'}^*}{\sum_g C_{gy} N_g^*} \quad (4.19)$$

where C_{gy} is the mean number of type- y contacts among members of patch g . If C_{gy} does not vary by patch, then (4.19) can be simplified by removing all C_{gy} terms. Equations (4.18)–(4.19) can then be repeated for each mobile pool g^* , reflecting the contributions of the *mobile-away* and *mobile-at-home* situations to overall mixing.

For *non-mobile-at-home* contacts, $M_{gg'}$ can effectively be replaced with an identity matrix $\delta_{gg'}$, which allows the following simplification:

$$X_{gg'y}^h = h_y C_{gy} N_g \delta_{gg'} \quad (4.20)$$

The total number of type- y contacts between patches g and g' can then be defined as the sum across all mixing pools:

$$X_{gg'y} = X_{gg'y}^h + \sum_{g^*} X_{gg'y}^* \quad (4.21)$$

Since $X_{gg'y}$ is defined at the population scale, we can verify that this matrix is symmetric and transform the result to per-person scale $C_{gg'y}$ or probability scale $\phi_{gg'y}$ if needed:

$$C_{gg'y} = X_{gg'y} N_g^{-1} \quad (4.22)$$

$$\phi_{gg'y} = C_{gg'y} C_g^{-1} \quad (4.23)$$

Additionally, if the mobility matrix $M_{gg'}$ changes with time, such changes can be propagated through the equations above to obtain a dynamic contact matrix.

4.6.3 Integrating Age Mixing and Mobility Data in Contact Matrices

Finally, we are ready to construct contact matrices which integrate the age-contact propensity matrix $\Gamma_{aa'y}$ from [23], and a given mobility matrix $M_{gg'}$. Here, for simplicity, we assume that contact numbers do not vary by patch, although it should be possible to extend the following approach to avoid this assumption.

We begin again by defining the numbers of individuals present in a given mobile pool g^* , adapting (4.18) to include age stratification:

$$N_{ga}^* = N_{ga} M_{gg^*} \quad (4.24)$$

The total number of individuals present from age group a is then given by the sum over all patches: $N_a^* = \sum_g N_{ga}^*$. Since mixing by patches gg' within pool g^* is assumed to be random, this mixing can be computed independently for each combination of age groups aa' . Thus, the proportion of contacts formed between patches g and g' , among the total contacts formed between age groups a and a' , can be defined as:

$$F_{gg',aa'}^* = \frac{N_{ga}^* N_{g'a'}^*}{N_a^* N_{a'}^*} \quad (4.25)$$

If contacts were to vary by patch, (4.25) should incorporate this information, similar to (4.19). Next, the age-contact propensity matrix for type- y contacts can be adapted to the population present in the mobile pool g^* and scaled by N_a^* to give the population-scale age mixing matrix:

$$X_{aa'y}^* = \Gamma_{aa'y} \frac{N_a^* N_a^*}{N^*} \quad (4.26)$$

Then, the total age-patch mixing within the pool is given by the multiplication of independent terms $X_{aa'y}^*$ and $F_{gg',aa'y}^*$:

$$X_{gag'a'y}^* = X_{aa'y}^* F_{gg',aa'y}^* \quad (4.27)$$

which conveniently simplifies to:

$$X_{gag'a'y}^* = \Gamma_{aa'y} \frac{N_{ga}^* N_{g'a'}^*}{N^*} \quad (4.28)$$

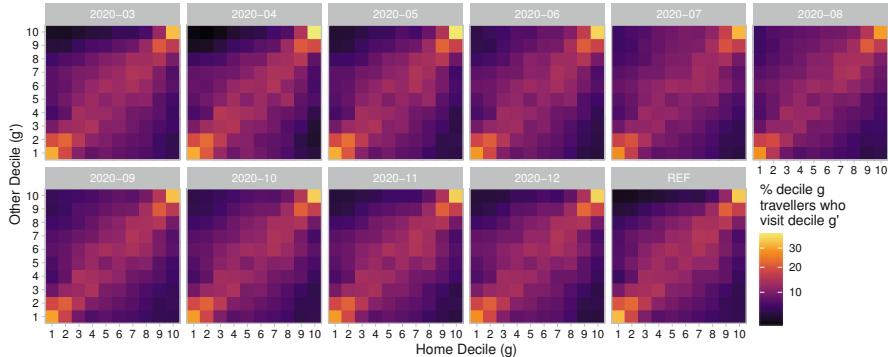
The contributions of each mixing pool to overall mixing can then be summed as before in (4.21):

$$X_{gag'a'y} = X_{gag'a'y}^h + \sum_{g^*} X_{gag'a'y}^* \quad (4.29)$$

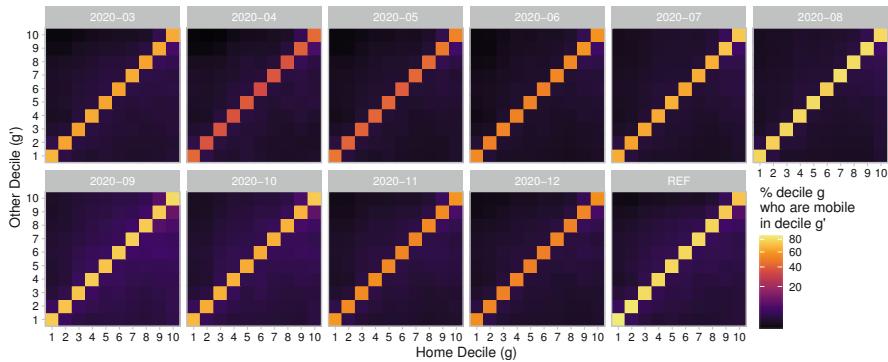
4.6.4 Example

Figure 4.9 illustrates mobility matrices for the ten Ontario patches in the motivating example. Additional details about the methods and assumptions used to generate these matrices are available in the appendix of [16]. Figure 4.9a illustrates the probability of travelling to any patch, among individuals who are *mobile-away*—i.e. already travelling outside the home patch. Figure 4.9b illustrates the overall probability of travelling to any patch, for the total population—i.e. including *mobile-away*, *mobile-at-home*, and *non-mobile* individuals.

The modest clustering along the diagonal in Fig. 4.9a suggests that “mobile-away” individuals are more likely to travel to a patch with similar COVID-19 incidence as their home patch, possibly due to geographic clustering of incidence deciles (Fig. 4.1). In Fig. 4.9b, the diagonal clustering is dramatically increased, due to the predominance of *mobile-at-home* individuals. This clustering then has important implications for the epidemic dynamics, as disproportionate contact of individuals at higher risk with other individuals at higher risk (i.e. assortativity) can accelerate initial epidemic spread but also concentrate and saturate transmissions within the higher-risk clusters to produce a lower endemic prevalence [10]. Such patterns can similarly influence the impact of interventions prioritized by patch.



(a) Conditional mobility matrix among travellers outside their home neighbourhood



(b) Absolute mobility matrix, including mobility within the home neighbourhood

Fig. 4.9 Mobility matrices for Ontario patches (deciles) during March–December 2020, plus the reference period (REF) of January–February 2020. Mobility derived from mobile device geolocation data [11]. Deciles represent groupings of Ontario neighbourhoods by cumulative COVID-19 cases, 15 January 2020–28 March 2021. Colour scale is square-root transformed to improve perception of smaller values

Another characteristic of Fig. 4.9b is that the probabilities only sum to 1 for the reference period; during the subsequent months (March–December 2022), observed reductions in mobility due to COVID-19 are explicitly incorporated into the mobility matrices. These reductions are then reflected in the downstream contact matrices via (4.24).

Following the methods in Sect. 4.6.3, the mobility matrices above were combined with the restratified age mixing matrices from Sect. 4.5.2, to obtain the fully stratified contact matrices we set out to obtain in Sect. 4.2. These contact matrices span 6 dimensions: 10×10 patches (self \times other), 12×12 age groups (self \times other), 4 contact types, and 11 time periods, for a total of 633,600 elements. As such, it can be difficult to visualize the matrices.

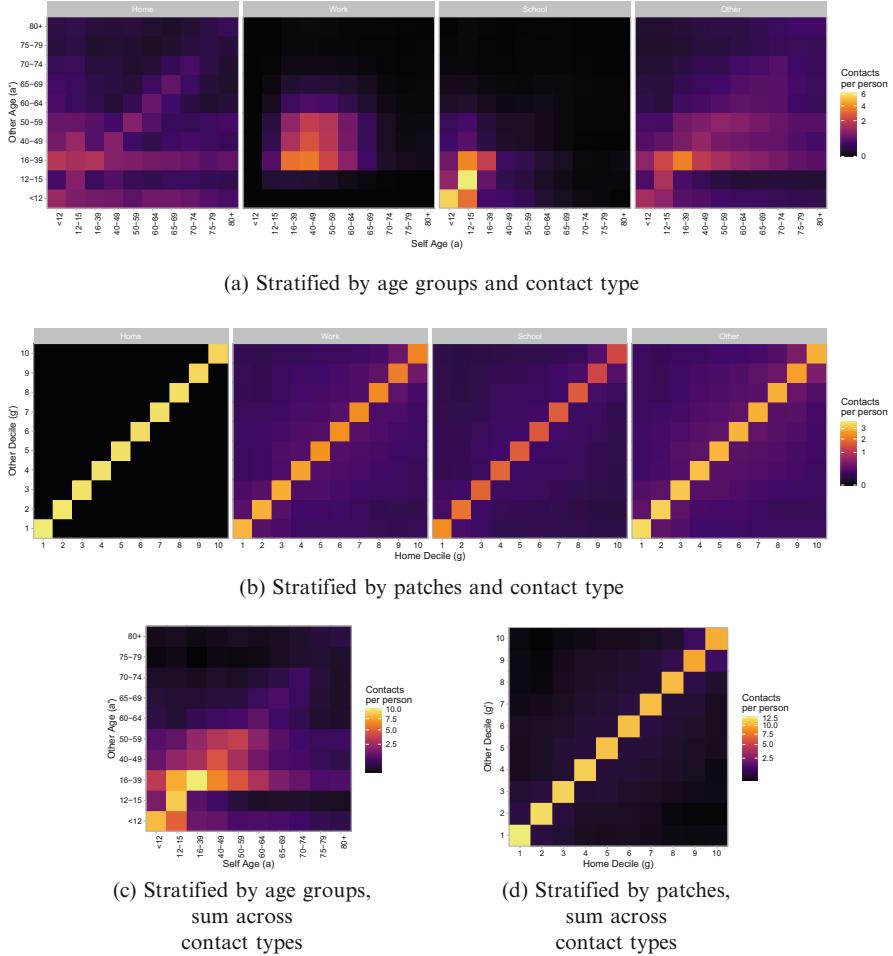


Fig. 4.10 Modelled contacts per person per day, stratified by age, decile (patch), and contact type, computed as the margins of the overall contact matrices $C_{gag'a'y}$. Contact matrices for Canada, derived using contact data from [23] and mobility data from [11] (January–February 2022). Deciles represent groupings of Ontario neighbourhoods by cumulative COVID-19 cases, 15 January 2020–28 March 2021. Colour scales are square-root transformed to improve perception of smaller values

Figure 4.10 summarizes the matrices on the per-person scale, by aggregating across different combinations of dimensions for the reference period only. As noted in Sect. 4.5.1, aggregation on the per-person scale entails summing across “other” dimensions and averaging across “self” dimensions. Figure 4.10a and c aggregate across the patch dimensions to obtain age-stratified matrices by contact type and overall, respectively. Figure 4.10b and d likewise aggregate across the age dimensions to obtain patch-stratified matrices by contact type and overall,

respectively. The age-stratified contact matrices are similar to the input matrices $\Gamma_{aa'y}$ from Fig. 4.3, only now weighted by the 2016 Ontario population distribution N_a . By contrast, the patch-stratified contact matrices are distinct from the mobility matrices in Fig. 4.9. Specifically, the contact matrices for work, school, and others include more mixing between patches than the mobility matrices alone would suggest, due to the case where *mobile-away* travellers from two patches meet in a third patch. Conversely, the home contact matrices feature no mixing between patches, since we assume that all contacts are formed with other residents of the same patch ($h_y = 1$).

Acknowledgments The study was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC CGS-D); the Canadian Institutes of Health Research (VR5-172683); and the 2020 COVID-19 Centred Research Award from the St Michael's Hospital Foundation Research Innovation Council.

From Unity Health Toronto, we thank Mackenzie Hamilton for helpful discussions and support in conceptualizing the chapter; Linwei Wang, Korryn Bodner, and Huiting Ma for helpful discussions; Kristy Yiu for research coordination support; and Gary Moloney for support with geographic data processing.

The data and R code used for the motivating example are available at github.com/mishra-lab/age-patch-mobility-mixing.

References

1. Adu, P.A., Binka, M., Mahmood, B., Jeong, D., Buller-Taylor, T., Damascene, M.J., Iyanwura, S., Ringa, N., García, H.A.V., Wong, S., Yu, A., Bartlett, S., Wilton, J., Irvine, M.A., Otterstatter, M., Janjua, N.Z., Naveed, D., Janjua, Z.: Cohort profile: the British Columbia COVID-19 Population Mixing Patterns Survey (BC-Mix). *BMJ Open* **12**(8), e056615 (2022). <https://doi.org/10.1136/BMJOOPEN-2021-056615>
2. Arenas, A., Cota, W., Gómez-Gardeñes, J., Gómez, S., Granell, C., Matamalas, J.T., Soriano-Paños, D., Steinegger, B.: Modeling the spatiotemporal epidemic spreading of COVID-19 and the impact of mobility and social distancing interventions. *Phys. Rev. X* **10**(4), 041055 (2020). <https://doi.org/10.1103/PhysRevX.10.041055>. <http://www.doi.org/10.1103/PhysRevX.10.041055>
3. Arregui, S., Aleta, A., Sanz, J., Moreno, Y.: Projecting social contact matrices to different demographic structures. *PLoS Comput. Biol.* **14**(12), e1006638 (2018). <https://doi.org/10.1371/journal.pcbi.1006638>
4. Baral, S., Beyerer, C., Muessig, K., Poteat, T., Wirtz, A.L., Decker, M.R., Sherman, S.G., Kerrigan, D.: Burden of HIV among female sex workers in low-income and middle-income countries: A systematic review and meta-analysis. *Lancet Infect. Dis.* **12**(7), 538—549 (2012). [https://doi.org/10.1016/S1473-3099\(12\)70066-X](https://doi.org/10.1016/S1473-3099(12)70066-X)
5. Begon, M., Bennett, M., Bowers, R.G., French, N.P., Hazel, S.M., Turner, J.: A clarification of transmission terms in host-microparasite models: Numbers, densities and areas. *Epidemiol. Infect.* **129**(1), 147–153 (2002). <https://doi.org/10.1017/S0950268802007148>
6. Béhanzin, L., Diabaté, S., Minani, I., Lowndes, C.M., Boily, M.C., Labbé, A.C., Anagonou, S., Zannou, D.M., Buvé, A., Alary, M.: Assessment of HIV-related risky behaviour: A comparative study of face-to-face interviews and polling booth surveys in the general population of Cotonou, Benin. *Sex. Transm. Infect.* **89**(7), 595–601 (2013). <https://doi.org/10.1136/sextrans-2012-050884>

7. Beyerer, C., Baral, S.D., Van Griensven, F., Goodreau, S.M., Chariyalertsak, S., Wirtz, A.L., Brookmeyer, R.: Global epidemiology of HIV infection in men who have sex with men. *Lancet* **380**(9839), 367–377 (2012). [https://doi.org/10.1016/S0140-6736\(12\)60821-6](https://doi.org/10.1016/S0140-6736(12)60821-6)
8. Brankston, G., Merkley, E., Fisman, D.N., Tuite, A.R., Poljak, Z., Loewen, P.J., Greer, A.L.: Quantifying contact patterns in response to COVID-19 public health measures in Canada. *BMC Public Health* **21**(1) (2021). <https://doi.org/10.1186/s12889-021-12080-1>
9. Brisson, M., Drolet, M., Mondor, M., Godbout, A., Gingras, G., Demers, É.: CONNECT: étude des contacts sociaux des Québécois (2021). <https://www.inspq.qc.ca/covid-19/donnees-connect/21-juillet-2021>
10. Garnett, G.P., Anderson, R.M.: Sexually transmitted diseases and sexual behavior: Insights from mathematical models. *J. Infect. Dis.* **174**(SUPPL. 2), S150–S161 (1996). https://doi.org/10.1093/infdis/174.supplement_2.s150
11. Ghasemi, A., Daneman, N., Berry, I., Buchan, S.A., Soucy, J.P., Sturrock, S., Brown, K.A.: Impact of a nighttime curfew on overnight mobility. *medRxiv* p. 2021.04.04.21254906 (2021). <https://doi.org/10.1101/2021.04.04.21254906>
12. Granell, C., Mucha, P.J.: Epidemic spreading in localized environments with recurrent mobility patterns. *Phys. Rev. E* **97**(5), 052302 (2018). <https://doi.org/10.1103/PhysRevE.97.052302>
13. Hu, T., Wang, S., She, B., Zhang, M., Huang, X., Cui, Y., Khuri, J., Hu, Y., Fu, X., Wang, X., Wang, P., Zhu, X., Bao, S., Guan, W., Li, Z.: Human mobility data in the COVID-19 pandemic: characteristics, applications, and challenges. *Int. J. Digital Earth* **14**(9), 1126–1147 (2021). <https://doi.org/10.1080/17538947.2021.1952324>
14. Johnstone-Robertson, S.P., Mark, D., Morrow, C., Middelkoop, K., Chiswell, M., Aquino, L.D., Bekker, L.G., Wood, R.: Social mixing patterns within a South African township community: Implications for respiratory disease transmission and control. *Am. J. Epidemiol.* **174**(11), 1246–1255 (2011). <https://doi.org/10.1093/aje/kwr251>
15. Kleynhans, J., Tempia, S., McMorrow, M.L., von Gottberg, A., Martinson, N.A., Kahn, K., Moyes, J., Mkhencelle, T., Lebina, L., Gómez-Olivé, F.X., Wafawana, F., Mathunjwa, A., Cohen, C., Buys, A., Mathee, A., Language, B., Maake, L., Treurnicht, F., Mothlaoleng, K., Carrim, M., Wolter, N., Hellfersree, O., Wagner, R.G., Piketh, S.: A cross-sectional study measuring contact patterns using diaries in an urban and a rural community in South Africa, 2018. *BMC Public Health* **21**(1), 1–10 (2021). <https://doi.org/10.1186/s12889-021-11136-6>
16. Knight, J., Ma, H., Ghasemi, A., Hamilton, M., Brown, K., Mishra, S.: Adaptive data-driven age and patch mixing in contact networks with recurrent mobility. *MethodsX* **9**, 101614 (2022). <https://doi.org/10.1016/j.mex.2021.101614>
17. Mishra, S., Stall, N.M., Ma, H., Odutayo, A., Kwong, J.C., Allen, U., Brown, K.A., Bogoch, I.I., Erman, A., Huynh, T., Ikura, S., Maltsev, A., McGeer, A., Moloney, G., Morris, A.M., Schull, M., Siddiqi, A., Smylie, J., Watts, T., Yiu, K., Sander, B., Jin, P.: A vaccination strategy for Ontario COVID-19 hotspots and essential workers. *Science Briefs of the Ontario COVID-19 Science Advisory Table*, vol. 2(26) (2021). <https://doi.org/10.47326/ocsat.2021.02.26.1.0>
18. Morris, M.: A log-linear modeling framework for selective mixing. *Math. Biosci.* **107**(2), 349–77 (1991). [https://doi.org/10.1016/0025-5564\(91\)90014-a](https://doi.org/10.1016/0025-5564(91)90014-a)
19. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., Edmunds, W.J.: Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**(3), 0381–0391 (2008). <https://doi.org/10.1371/journal.pmed.0050074>
20. Nold, A.: Heterogeneity in disease-transmission modeling. *Math. Biosci.* **52**(3-4), 227–240 (1980). [https://doi.org/10.1016/0025-5564\(80\)90069-3](https://doi.org/10.1016/0025-5564(80)90069-3)
21. Nyquist, H.: Certain topics in telegraph transmission theory. *Trans. Am. Instit. Elect. Eng.* **47**(2), 617–644 (1928). <https://doi.org/10.1109/T-AIEE.1928.5055024>
22. Prem, K., Cook, A.R., Jit, M.: Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS Comput. Biol.* **13**(9), e1005697 (2017). <https://doi.org/10.1371/journal.pcbi.1005697>

23. Prem, K., van Zandvoort, K., Klepac, P., Eggo, R.M., Davies, N.G., Cook, A.R., Jit, M.: Projecting contact matrices in 177 geographical regions: An update and comparison with empirical data for the COVID-19 era. *PLOS Comput. Biol.* **17**(7) (2021). <https://doi.org/10.1371/journal.pcbi.1009098>
24. Ruschendorf, L.: Convergence of the iterative proportional fitting procedure. *Ann. Stat.* **23**(4), 1160–1174 (1995). <https://doi.org/10.1214/aos/1176324703>
25. Simini, F., González, M.C., Maritan, A., Barabási, A.L.: A universal model for mobility and migration patterns. *Nature* **484**(7392), 96–100 (2012). <https://doi.org/10.1038/nature10856>

Chapter 5

An Optimal Control Approach for Public Health Interventions on an Epidemic-Viral Model in Deterministic and Stochastic Environments



Idriss Sekkak and Bouchra R. Nasri

5.1 Introduction

The most challenging part of making decisions is not providing the correct action, but it is having to take predictive measures on that decision, determining the risk level, and building a dashboard that tracks the consequences. In times of a pandemic, there are two key mechanisms that worry a public health decision-maker: the first one is the dynamics of an epidemic and the infectious disease spread in a population, and the second one is the immunological response and the viral load of an individual contributing in the disease transmission. In the literature, viral models [1, 16–20] take into account the dynamics inside the host that are unrelated to the contact between individuals at the macro level. On the other hand, epidemic models [3, 5, 6, 12, 22] based on the basic model of Kermack and McKendrick [8] investigate the interlinkage between susceptible and infected individuals without considering the effect of viral load of each individual.

Nevertheless, there are some problems that can only be investigated using frameworks that formally connect the two systems in order to explore some challenges such as the following:

- How does the viral load of one individual impact his surroundings?
- Will a public health intervention impact both dynamics?
- What is the repercussion of public health interventions on coupling both models and their basic reproduction numbers for viral and epidemic models?
- Will an environmental perturbation impact the linked dynamics?

I. Sekkak (✉) · B. R. Nasri

Centre de recherche en santé publique and Département de médecine sociale et préventive, École de santé publique de l'Université de Montréal, Montréal, QC, Canada
e-mail: bouchra.nasri@umontreal.ca

Therefore, this work seeks to investigate those questions based on the proposed framework by Feng et al. [7], where authors separated the biological time scales for the coupled dynamics, a fast time scale related to the viral dynamics, and a slow time scale associated with the epidemiological and environmental dynamics.

5.1.1 A Fast Time Scale Viral Model

In order to understand the dynamics of viruses, authors in [14] drove a basic idea to analyze the HIV infection kinetics, and that resulted in the development of a very known field called viral dynamics. First, the patient with a viral disease reaches a set-point level of viral load and then remains at approximately that level for an amount of time depending on the viral properties. Hence, to maintain this equilibrium level, the individual's body must clear the virus at the same rate it is generated. A viral model can be described as

$$\begin{aligned} dT &= [\Lambda - kVT - mT] dt, \\ dT^* &= [kVT - (m + d)T^*] dt, \\ dV &= [g(E) + pT^* - cV] dt, \end{aligned} \quad (5.1)$$

where the compartments T , T^* , and V represent, respectively, the healthy cells, infected cells, and free virions. The parameters are described as follows: Λ denotes the generation rate of healthy cells that decay naturally at the rate m , and they get infected by the attacks of free virions produced from the infected cells with a transmission rate described by k . The infected cells decay at a rate of $m + d$, where d represents the mortality rate of infected cells rate related to the infection. The virus reproduction rate by an infected cell is described by p . The death rate of free virions is c . In addition, the function $g(E)$ describes the rate at which an average host is infused by an outsider viral load, where $E = E(t)$ determines the level of a contaminated environment at time t , with areas being considered as $0 \leq E \leq 1$, such that

$$dE = [\theta IV(1 - E) - \gamma E] dt. \quad (5.2)$$

The degree of contamination is dependent on the average viral load V within the host and the number of infected hosts I , where θ describes the contamination rate and γ is the decay rate in a contaminated environment. For biological meaning, it is considered that the function g should have the following assumptions:

$$g(E) \geq 0, \quad g(0) = 0, \quad g'(E) > 0. \quad (5.3)$$

Besides, a life of a virus in an outside body environment is known to be challenging for this pathogen, but it can be described as a slower time scale than the cycle of a virus in an individual's body since an immune response can impact its existence.

5.1.2 SIQR Epidemic Model with a Coupled Viral Model

The most important aspects of an infectious disease propagation in a population can be described using the SIR model [8], which is one of the main approaches used to predict an epidemic behavior, and several models in the literature are derivatives of this basic form, such as the use of a quarantine strategy [3, 5], the loss of immunity [11], or the incorporation of media coverage [2] to slow down the spread of the disease. The epidemic model consists of three compartments: S susceptible individuals to get infected, I infected individuals, and R recovered or immune individuals. In this work, based on the incorporation of within-host dynamics into a SI epidemic model in the work of Feng et al. [7], we include the viral system (5.1) and contamination area (5.2) into a SIQR epidemic model.

$$\begin{aligned} dS &= [\mu N - \beta E(t)S(t) - \mu S(t)] dt, \\ dI &= [\beta E(t)S(t) - (\mu + \lambda_1 + \lambda_2 + \alpha_1)I(t)] dt, \\ dQ &= [\lambda_1 I(t) - (\varepsilon + \mu + \alpha_2)Q(t)] dt, \\ dR &= [\lambda_2 I(t) + \varepsilon Q(t) - \mu R(t)] dt, \end{aligned} \quad (5.4)$$

where $N = S + I + Q + R$. The parameters are described as follows: μ denotes the recruitment and death rate, β describes the transmission rate, λ_1 denotes the quarantine rate of infected individuals, λ_2 and ε represent the recovery rate for infected and quarantined individuals, respectively, while α_1 and α_2 denote the death rate related to the infectious disease. However, the notion of time differs from virus infection and its generation to the spread of the disease in a population or an environment.

5.1.3 Qualitative Analysis of the Coupled Model

An important biological fact of this coupled framework dynamics is that the viral dynamics happen on a faster time scale than the dynamics of the epidemic and the environmental dynamics. This difference in time progression motivated Feng et al. [7] to investigate the behavior properties of models (5.1), (5.2), and (5.7) by exploring separately the faster and slower time scale dynamics. First, we handle some proprieties of the fast time scale viral model (5.1). Therefore, if we consider the nonexistence of contaminated area denoted as $g(E) = 0$, then the system will

have two equilibrium points, the free infection state $W_0 = (T_0, T_0^*, V_0) = (\frac{\Lambda}{m}, 0, 0)$ and nontrivial infection equilibrium $W^* = (\tilde{T}, \tilde{T}^*, \tilde{V})$. However, in case of $g(E) > 0$, the system has only an infection state. We denote the basic reproduction number as $\mathcal{R}_v(E)$, where the initial basic within-host reproduction number can be computed using the next-generation method as in [4]

$$\mathcal{R}_v(0) = \mathcal{R}_{v0} = \frac{T_0 kp}{c(m + d)}. \quad (5.5)$$

According to [7], if we assume that $\mathcal{R}_v > 1$, it results that the fast within-host system is at the stable nontrivial equilibrium $\tilde{W}_-(E) = (\tilde{T}_-(E), \tilde{T}_-^*(E), \tilde{V}_-(E))$, such that

$$\begin{aligned} \tilde{T}_-(E) &= \frac{T_0}{\mathcal{R}_v(E)}, & \tilde{V}_-(E) &= \frac{1}{c} \left[g(E) + \frac{p\Lambda}{m+d} \left(1 - \frac{1}{\mathcal{R}_v(E)} \right) \right], \\ \tilde{T}_-^*(E) &= \frac{p\Lambda}{m+d} \left(1 - \frac{1}{\mathcal{R}_v(E)} \right). \end{aligned}$$

Hence, the basic reproduction number can be formulated as

$$\mathcal{R}_v(E) =: \frac{T_0}{\tilde{T}_-(E)}. \quad (5.6)$$

In this work, we consider the epidemiological and environmental variables to be slow variables, which consist of S, I, Q, R , and E . Hence, we include the fast time scale variable at their equilibrium state, since the infection within a human body generates a viral load that can contaminate the surroundings of the infected individual. Therefore, we will incorporate the recovery and quarantine proprieties to investigate the following model:

$$\begin{aligned} dS &= [\mu N - \beta E(t)S(t) - \mu S(t)] dt, \\ dI &= [\beta E(t)S(t) - (\mu + \lambda_1 + \lambda_2 + \alpha_1)I(t)] dt, \\ dQ &= [\lambda_1 I(t) - (\varepsilon + \mu + \alpha_2)Q(t)] dt, \\ dR &= [\lambda_2 I(t) + \varepsilon Q(t) - \mu R(t)] dt, \\ dE &= \left[\theta I \tilde{V}_-(E)(1 - E(t)) - \gamma E(t) \right] dt. \end{aligned} \quad (5.7)$$

Applying the next-generation method as in [4], the baseline reproduction number for the coupled within-between-host model is formulated as

$$\mathcal{R}_{h0} =: \frac{\beta\theta N}{\gamma(\mu + \lambda_1 + \lambda_2 + \alpha_1)}. \quad (5.8)$$

Therefore, the reproduction number [7] for the coupled system is described as

$$\mathcal{R}_h =: \frac{pmT_0}{c(m+d)} \left(1 - \frac{1}{\mathcal{R}_{v0}}\right) \mathcal{R}_{h0}. \quad (5.9)$$

Throughout this work, we are interested in the investigation of the endemic case where $\mathcal{R}_{v0} > 1$ and $\mathcal{R}_h > 1$. Hence, we are exploring several strategies to handle a pandemic outbreak in order to ease its impact on the public health facilities. First, we consider model (5.7) and replace the deterministic transmission rate by a stochastic one in order to incorporate the incertitude of the environmental elements in the real world, which impacts the propagation of a disease in a population (such as humidity, wind, sunlight). This is done by letting

$$d\tilde{\beta}(t) = \beta dt + \sigma dW(t), \quad (5.10)$$

where σ denotes the intensity of the Brownian motion $W(t)$, which is determined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with filtration $\{\mathcal{F}_t\}_{t \geq 0}$ satisfying the usual conditions (i.e., it is right continuous and increasing, while \mathcal{F}_0 contains all \mathbb{P} -null sets), and we define the space \mathbb{R}_+^n as

$$\mathbb{R}_+^n = \{x \in \mathbb{R}^n : x_i > 0, \quad i = 1, \dots, n\}.$$

The resulting stochastic model is then

$$\begin{aligned} dS &= [\mu N - \beta E(t)S(t) - \mu S(t)]dt - \sigma E(t)S(t)dW(t), \\ dI &= [\beta E(t)S(t) - (\mu + \lambda_1 + \lambda_2 + \alpha_1)I(t)]dt + \sigma E(t)S(t)dW(t), \\ dQ &= [\lambda_1 I(t) - (\varepsilon + \mu + \alpha_2)Q(t)]dt, \\ dR &= [\lambda_2 I(t) + \varepsilon Q(t) - \mu R(t)]dt, \\ dE &= \left[\theta I \tilde{V}_-(E)(1 - E(t)) - \gamma E(t) \right]dt. \end{aligned} \quad (5.11)$$

For this matter, we handle a theorem on the existence and positivity of the global solution of the stochastic system (5.11) for any initial condition on \mathbb{R}_+^5 .

Theorem 1 *For any given initial value $(S(0), I(0), Q(0), R(0), E(0)) \in \mathbb{R}_+^5$, there exists a unique local solution $(S(t), I(t), Q(t), R(t), E(t))$ to system (5.11) on $t \geq 0$, and the solution remains in \mathbb{R}_+^5 almost surely.*

Proof Since the stochastic system (5.11) has locally Lipschitz coefficients, then, for any initial value $(S(0), I(0), Q(0), R(0), E(0)) \in \mathbb{R}_+^5$, there exists a unique local solution $(S(t), I(t), Q(t), R(t), E(t)) \in \mathbb{R}_+^5$ on $t \in [0, \tau_e]$, where τ_e denotes the explosion time. In order to get the global positivity of the solution, we need to prove that $\tau_e = \infty$ a.s. Define $k_0 > 0$ to be large enough so that $S(0), I(0), Q(0), R(0)$, and $E(0)$ belong to the interval $\left[\frac{1}{k_0}, k_0\right]$. In this matter, for each integer $k \geq k_0$, we consider the following stopping time:

$$\begin{aligned} \tau_k = \inf \left\{ t \in [0, \tau_e) : S(t) \notin \left(\frac{1}{k}, k \right) \text{ or } I(t) \notin \left(\frac{1}{k}, k \right) \text{ or } Q(t) \notin \left(\frac{1}{k}, k \right) \right. \\ \left. \text{or } R(t) \notin \left(\frac{1}{k}, k \right) \text{ or } E(t) \notin \left(\frac{1}{k}, k \right) \right\}, \end{aligned}$$

where τ_k is increasing as $k \uparrow \infty$. Set $\tau_\infty = \lim_{k \rightarrow \infty} \tau_k$. Therefore, $\tau_\infty \leq \tau_e$ a.s. Proving that $\tau_\infty = \infty$ means that $\tau_e = \infty$ and $(S(t), I(t), Q(t), R(t), E(t)) \in \mathbb{R}_+^5$ a.s. If this statement is false, then there exists a pair of constants $T > 0$ and $\varepsilon \in (0, 1)$ such that $\mathbb{P}\{\tau_\infty \leq T\} > \varepsilon$. Thus, there is an integer $k_1 \geq k_0$ such that

$$P\{\tau_k \leq T\} \geq \varepsilon, \quad \forall k \geq k_1. \quad (5.12)$$

Consider the C^2 -function $V_1 : \mathbb{R}_+^5 \rightarrow \mathbb{R}_+$ as follows:

$$\begin{aligned} V_1(S, I, Q, R, E) = & (S - 1 - \log S) + (I - 1 - \log I) + (Q - 1 - \log Q) \\ & + (R - 1 - \log R) + (E - 1 - \log E). \end{aligned}$$

Applying Itô's formula, we obtain

$$\begin{aligned} \mathcal{L}V_1 = & \left(1 - \frac{1}{S(t)} \right) (\mu N - \beta E(t)S(t) - \mu S(t)) + \frac{\sigma^2 E^2}{2} + \frac{\sigma^2 S^2 E^2}{I^2} \\ & + \left(1 - \frac{1}{I(t)} \right) (\beta E(t)S(t) - (\mu + \lambda_1 + \lambda_2 + \alpha_1)I(t)) \\ & + \left(1 - \frac{1}{Q(t)} \right) (\lambda_1 I(t) - (\varepsilon + \mu + \alpha_2)Q(t)) \\ & + \left(1 - \frac{1}{R(t)} \right) (\lambda_2 I(t) + \varepsilon Q(t) - \mu R(t)) \\ & + \left(1 - \frac{1}{E} \right) (\theta I \tilde{V}_-(E)(1 - E(t)) - \gamma E(t)) \\ \leq & \mu N + \beta + 3\mu + \lambda + \gamma + \frac{\sigma^2}{2} + \frac{\sigma^2 N^2}{I^2} \\ \leq & K. \end{aligned} \quad (5.13)$$

Integrating both sides from 0 to $\tau \wedge T$ leads to

$$\begin{aligned} \int_0^{\tau \wedge T} dV_1(S(s), I(s), Q(s), R(s), E(s)) \leq & \int_0^{\tau \wedge T} K ds \\ & + \sigma \int_0^{\tau \wedge T} S(s)E(s) \left(\frac{1}{S(s)} - \frac{1}{I(s)} \right) dW(s). \end{aligned} \quad (5.14)$$

Taking expectation of both sides of (5.14), we obtain

$$\begin{aligned} \mathbb{E}V_1(S(\tau \wedge T), I(\tau \wedge T), Q(\tau \wedge T), R(\tau \wedge T), E(\tau \wedge T)) \\ \leq V_1(S(0), I(0), Q(0), R(0), E(0)) + KT. \end{aligned}$$

This yields to

$$V_1(S(\tau \wedge T), I(\tau \wedge T), Q(\tau \wedge T), R(\tau \wedge T), E(\tau \wedge T)) + KT \leq \varepsilon\theta_k, \quad (5.15)$$

where $\theta_k = (k - 1 - \log k) \wedge \left(\frac{1}{k} - 1 + \log k\right)$. Letting $k \rightarrow \infty$ yields to the contradiction $\infty > V_1(S(0), I(0), Q(0), R(0), E(0)) + KT = \infty$. This finishes the proof.

5.2 Optimal Control Analysis

In the following section, the optimal control problem to minimize an objective functional $\mathcal{J}(u)$ will be investigated. Therefore, we reconstruct the system (5.11) with the following controls:

$$\begin{aligned} dS &= [\mu N - \beta(1 - u_1(t))E(t)S(t) - (\mu + u_2(t))S(t)]dt, \\ dI &= [\beta(1 - u_1(t))E(t)S(t) - (\mu + u_3(t) + \lambda_1 + \lambda_2 + \alpha_1)I(t)]dt, \quad (5.16) \\ dQ &= [\lambda_1 I(t) - (\varepsilon + \mu + \alpha_1)Q(t)]dt, \\ dR &= [(\lambda_2 + u_3(t))I(t) + \varepsilon Q(t) + u_2(t)S(t) - \mu R(t)]dt, \\ dE &= \left[\theta I \tilde{V}_-(E)(1 - E(t)) - \gamma E(t) \right]dt, \end{aligned}$$

with $u_i(t) \in [0, u_{i_{max}}]$, and positive constants $u_i^{max} \leq 1$ for $i = \{1, 2, 3\}$; the controls $u_1(t)$, $u_2(t)$, and $u_3(t)$ refer to non-pharmaceutical interventions, vaccination campaigns, and treatment effort, respectively.

5.2.1 Investigation of the Deterministic Optimal Control

We consider an optimal control problem for minimizing the cost of the objective functional. In order to minimize the cost of the objective functional \mathcal{J} , we explore the following optimal control issue:

$$\begin{aligned} \mathcal{J}(u_1(t), u_2(t), u_3(t)) &= \int_0^T \left[A_1 I(t) + \frac{C_1 u_1^2(t)}{2} + \frac{C_2 u_2^2(t)}{2} + \frac{C_3 u_3^2(t)}{2} \right] dt, \\ (u_1, u_2, u_3) &\in U, \end{aligned} \quad (5.17)$$

where A_1 , C_1 , C_2 , and C_3 are positive weights and $S(0) > 0$, $I(0) > 0$, $Q(0) > 0$, $R(0) > 0$, and $E(0) > 0$, the term $\frac{C_1 u_1^2}{2}$ is related to contaminated environment control and non-pharmaceutical interventions u_1 , $\frac{C_2 u_2^2}{2}$ describes the vaccination strategy u_2 , and $\frac{C_3 u_3^2}{2}$ is related to the infection treatment control u_3 . Let $u = (u_1, u_2, u_3)$. In the following, we look for an optimal control $u^*(t)$ defined by

$$\mathcal{J}(u^*(t)) = \min_{u \in U} \{\mathcal{J}(u)\},$$

where the set of controls are defined as

$$U = \{u(t) = (u_i(t)) | 0 \leq u_i(t) \leq u_i^{max}; \forall t \in [0, T] \text{ and } i = \{1, 2, 3\}\}, \quad (5.18)$$

where the controls $u_i(t)$ are Lebesgue measurable. Besides, we recall Pontryagin's maximum principle [9] to define the formulation of our optimal controls u_i . We determine the Lagrangian by

$$L = A_1 I(t) + \frac{C_1 u_1^2(t)}{2} + \frac{C_2 u_2^2(t)}{2} + \frac{C_3 u_3^2(t)}{2}. \quad (5.19)$$

In order to determine the minimal value of the Lagrangian, we construct the following Hamiltonian function:

$$H = L(I, u) + p_1 \frac{dS}{dt} + p_2 \frac{dI}{dt} + p_3 \frac{dQ}{dt} + p_4 \frac{dR}{dt} + p_5 \frac{dE}{dt}. \quad (5.20)$$

Theorem 2 *There exists an optimal control u^* such that*

$$\mathcal{J}(u^*(t)) = \mathcal{J}(u(t)).$$

subject to the control system (5.16) with initial conditions.

Proof We consider the control and state variables with positive values and let the convexity condition be satisfied by the objective functional on the control set U . Therefore, the controls are convex and closed by definition. Besides, the boundedness of the optimal system determines the compactness required for the existence of the optimal control. Additionally, the integrand in the objective functional is convex on the control set U . Finally, there exist a constant $\rho > 1$,

$\omega_1 = A_2$, and $\omega_2 > 0$ such that

$$J(u_1(t)) \geq \omega_2 + \omega_1 \| u \|^\rho.$$

As result, we conclude from lemma [13] that there exists an optimal control solution u^* .

Next, we investigate an optimal solution pair by applying the Pontryagin's maximum principle to the Hamiltonian (5.20) for the optimal control problem. If $(x^*(t), u^*(t))$ is an optimal solution of an optimal control problem, then there exists nontrivial vector function $p = (p_1, p_2, p_3, p_4, p_5)$, where

$$\frac{\partial H}{\partial u_i} = 0 \quad \text{and } \dot{p} = \frac{p_i}{dt} = -\frac{\partial H}{\partial x_i}, \quad (5.21)$$

where $(x_1 = S; x_2 = I; x_3 = Q; x_4 = R; x_5 = E)$.

Theorem 3 Let $(S^*(t), I^*(t), Q^*(t), R^*(t), E^*(t))$ be optimal state solutions with associated optimal control variable $(u^*(t))$ for the optimal control problems (5.16) and (5.17). Then, there exist adjoint variables $(p_1, p_2, p_3, p_4, p_5)$ that satisfy the system (5.23) with transversality conditions $p_i(T) = 0$, for $i = 1, 2, 3, 4, 5$, where the optimal control $u^*(t)$ is

$$u_1^* = \frac{(p_2 - p_1)\beta E^* S^*}{C_1}, \quad u_2^* = \frac{(p_1 - p_4)S^*}{C_2} \quad u_3^* = \frac{(p_2 - p_4)I^*}{C_3} \quad (5.22)$$

Proof First, we define $p_i(t)$, $i = 1, 2, 3, 4, 5$ by differentiating the Hamiltonian (5.20) with respect to the state variables (S, I, Q, R, E) . Hence, we obtain

$$\begin{aligned} \frac{dp_1}{dt} &= -\frac{\partial H}{\partial S} = p_1(\beta(1 - u_1(t))E + u_2(t) + \mu) - p_2\beta(1 - u_1(t))E - p_4u_2(t), \\ \frac{dp_2}{dt} &= -\frac{\partial H}{\partial I} = -A_1 + p_2(\mu + u_3(t) + \lambda_1 + \lambda_2 + \alpha_1) - p_3\lambda_1 - p_4(\lambda_2 + u_3(t)) \\ &\quad - p_5\theta\tilde{V}_-(E)(1 - E), \\ \frac{dp_3}{dt} &= -\frac{\partial H}{\partial Q} = p_3(\varepsilon + \mu + \alpha_2) - p_4\varepsilon, \\ \frac{dp_4}{dt} &= -\frac{\partial H}{\partial R} = p_4\mu, \\ \frac{dp_5}{dt} &= -\frac{\partial H}{\partial E} = p_1\beta(1 - u_1(t))S - p_2\beta(1 - u_1(t))S \\ &\quad - p_5\left(\theta I \frac{d\tilde{V}_-}{dE} - \theta IE \frac{d\tilde{V}_-}{dE}\right) + p_5\left(\theta I\tilde{V}_- + \gamma\right), \end{aligned} \quad (5.23)$$

where

$$\frac{d\tilde{V}_-}{dE} = \frac{g'(E)}{c} - \frac{p\Lambda}{m+d} \frac{\tilde{T}'_-(E)}{T_0}. \quad (5.24)$$

By considering the transversality conditions $p_i(T) = 0$, for $i = 1, 2, 3, 4, 5$, and according to Pontryagin's maximum principle, we determine the optimal conditions as

$$\frac{\partial H}{\partial u_1} = C_1 u_1(t) - (p_2 - p_1)\beta E S = 0,$$

$$\frac{\partial H}{\partial u_2} = C_2 u_2(t) - (p_1 - p_4)S(t) = 0,$$

$$\frac{\partial H}{\partial u_3} = C_3 u_3(t) - (p_2 - p_4)I(t) = 0,$$

Hence, we obtain (5.22), with the following properties of the control set

$$u_1^m = \begin{cases} 0 & \text{if } p_2 - p_1 < 0, \\ \frac{(p_2 - p_1)\beta E^* S^*}{C_1} & \text{if } 0 \leq (p_2 - p_1)\beta E^* S^* \leq C_1, \\ u_1^{max} & \text{if } (p_2 - p_1)\beta E^* S^* \geq C_1, \end{cases}$$

$$u_2^m = \begin{cases} 0 & \text{if } p_1 - p_4 < 0, \\ \frac{(p_1 - p_4)S^*}{C_2} & \text{if } 0 \leq (p_1 - p_4)S^* \leq C_2, \\ u_2^{max} & \text{if } (p_1 - p_4)S^* \geq C_2, \end{cases}$$

$$u_3^m = \begin{cases} 0 & \text{if } p_2 - p_4 < 0, \\ \frac{(p_2 - p_4)I^*}{C_3} & \text{if } 0 \leq (p_2 - p_4)I^* \leq C_3, \\ u_3^{max} & \text{if } (p_2 - p_4)I^* \geq C_3. \end{cases}$$

5.2.2 Investigation of the Stochastic Optimal Control

In the following part, we explore the stochastic optimization problem and define its solution.

$$dS = [\mu N - \beta(1 - u_1(t))E(t)S(t) - u_2(t)S(t)]dt - \sigma E(t)S(t)dW(t),$$

$$\begin{aligned}
dI &= [\beta(1 - u_1(t))E(t)S(t) - (\mu + u_3(t) + \lambda_1 + \lambda_2 + \alpha_1)I(t)]dt \\
&\quad + \sigma E(t)S(t)dW(t), \\
dQ &= [\lambda_1 I(t) - (\varepsilon + \mu + \alpha_2)Q(t)]dt, \\
dR &= [(\lambda_2 + u_3(t))I(t) + \varepsilon Q(t) - \mu R(t) + u_2(t)S(t)]dt, \\
dE &= \left[\theta I \tilde{V}_-(E)(1 - E(t)) - \gamma E(t) \right]dt,
\end{aligned} \tag{5.25}$$

where our motivation is to investigate optimal non-pharmaceutical interventions, vaccination policy, and treatment rates $u_i^*(t)$ $i = \{1, 2, 3\}$ and to determine an optimal effort on the infectious size that minimizes the objective functional for which an initial state x_0 is determined as

$$E_{0,x_0} \left[\int_0^T \left(A_1 I(t) + \frac{C_1 u_1^2(t)}{2} + \frac{C_2 u_2^2(t)}{2} + \frac{C_3 u_3^2(t)}{2} \right) ds \right]. \tag{5.26}$$

We assume that there is positive constant $u_i^m \leq 1$ such that $u_i(t) \leq u_i^{max}$ a.s. and the class of admissible control law is

$$\mathcal{A} = \{u(\cdot) : u \text{ are adapted} \mid 0 \leq u(t) \leq u^{max} \text{ a.s.}\}.$$

Therefore, to find a solution to this stochastic problem, we determine the performance criterion of the form as

$$\mathcal{J}(t, x, u_1) = E_{t,x} \left[\int_0^T \left(A_1 I(t) + \frac{C_1 u_1^2(t)}{2} + \frac{C_2 u_2^2(t)}{2} + \frac{C_3 u_3^2(t)}{2} \right) ds \right], \tag{5.27}$$

where the expectation is conditional on the state of the system being a fixed value at x at time t . Next, we define the value function and the optimal control $u^* \in \mathcal{A}$ as

$$\Phi(t, x) = \inf_{u(\cdot) \in \mathcal{A}} \mathcal{J}(t, x, u) = \mathcal{J}_s(t, x, u^*), \tag{5.28}$$

where we determine a control law that minimizes the value $\mathcal{J} : \mathcal{A} \rightarrow \mathbb{R}_+$ given by (5.27). Hence, to find a solution to our stochastic optimal control problem for the stochastic system (5.25) given \mathcal{A} in (5.28) and \mathcal{J} in (5.27), we define the value of the function Φ and a control function

$$u^* = \arg \inf_{u(t) \in \mathcal{A}} \mathcal{J}_s(x; u(t)) \in \mathcal{A}. \tag{5.29}$$

For this matter, we show the existence of optimal control under the stochastic formulation, where we have the term of controllability on the drift part. We recall that the results of existence were investigated by Kushner [10].

Theorem 4 *There exists a stochastic optimal control u^* such that*

$$\mathcal{J}_s(u^*(t)) = \inf_{u \in \mathcal{A}} \mathcal{J}(u(t))$$

subject to the control system (5.25) with initial conditions.

Proof According to Yong and Zhou [21], the existence of optimal control is fulfilled if (U, d) is a Polish space and $T > 0$ and the maps F, G, L are uniformly continuous and there exists a constant $L > 0$ such that $\phi(t, x, u) = F(t, x, u), G(t, x), \phi(x_u(T))$

$$\begin{aligned} |\phi(t, x, u) - \phi(t, \hat{x}, u)| &\leq L|x - \hat{x}|, \quad \forall t \in [0, T], x, \hat{x} \in \mathbb{R}^n, u \in U, \\ |\phi(t, 0, u)| &\leq L, \quad \forall (t, u) \in [0, T] \times U. \end{aligned}$$

In the following theorem, we express the characterization of stochastic optimal control $u_1^*(t)$.

Theorem 5 *A solution to the optimal non-pharmaceutical interventions, vaccination, and treatment strategies problems stated in problem (5.28) is of the form*

$$\begin{aligned} u_1^* &= \min \left\{ \max \left\{ 0, \frac{(\Phi_I - \Phi_S)\beta E^* S^*}{C_1} \right\}, u_1^{max} \right\}, \\ u_2^* &= \min \left\{ \max \left\{ 0, \frac{(\Phi_S - \Phi_R)S^*}{C_2} \right\}, u_2^{max} \right\}, \\ u_3^* &= \min \left\{ \max \left\{ 0, \frac{(\Phi_I - \Phi_R)I^*}{C_3} \right\}, u_3^{max} \right\}. \end{aligned} \quad (5.30)$$

Proof By applying Itô's formula, we obtain

$$\begin{aligned} \mathcal{L}\Phi &= \dot{S}\Phi_S(t) + \dot{I}\Phi_I(t) + \dot{Q}\Phi_Q(t) + \dot{R}\Phi_R(t) + \dot{E}\Phi_E(t) + \frac{(\sigma S(t)E(t))^2}{2}\Phi_{SS}(t) \\ &\quad + \frac{(\sigma S(t)E(t))^2}{2}\Phi_{II}(t) - \frac{(\sigma S(t)E(t))^2}{2}\Phi_{SS}(t) - \frac{(\sigma S(t)E(t))^2}{2}\Phi_{II}(t). \end{aligned}$$

Additionally, we apply Hamiltonian Jacobi Bellman theory [15] to get the following infinitum of (5.28):

$$\inf_{u(\cdot) \in \mathcal{A}} \left[A_1 I(t) + \frac{C_1 u_1^2(t)}{2} + \frac{C_2 u_2^2(t)}{2} + \frac{C_3 u_3^2(t)}{2} + \mathcal{L}\Phi(t) \right]. \quad (5.31)$$

Therefore, by deriving (5.31) with respect to $u(t)$, we get

$$\begin{aligned} C_1 u_1(t) - (\Phi_I - \Phi_S)\beta E S &= 0, \\ C_2 u_2(t) - (\Phi_S - \Phi_R)S(t) &= 0, \\ C_3 u_3(t) - (\Phi_I - \Phi_R)I(t) &= 0. \end{aligned}$$

Finally, by considering the bounds on $(u_i(t))$ and by the same argument in the deterministic characterization, we get the asserted expression (5.30).

5.3 Numerical Simulations

In this section, we illustrate numerical scenarios for the deterministic stochastic (5.11) coupled models with and without timely control in order to investigate numerically the role of treatment to control the spread of an infectious disease in a population. Therefore, we set a basic reproduction number leading to an endemic case $R_0 > 1$ to explore how the patterns will behave using an optimal control approach. Using the values in the Table 5.1 and for initial states values as $(S, I, Q, R, E) = (200,000, 50, 10, 200, 0.1)$, we construct a simulation with a basic reproduction number $R_0 = 1.3638$.

As result, in Fig. 5.1 we can see the impact of the controls on the behavior to minimize the size of the infected population. In Fig. 5.6, we illustrate a timely control strategy using non-pharmaceutical interventions u_1 , where according to simulation it is recommended to start with strong measures to control the spread in the first 4 days followed by relaxation, while a vaccination policy is increasing gradually at a slow rate to support the NPI measures in time of relaxation. Meanwhile, the treatment strategy in Fig. 5.7 follows the behavior of NPI measures to minimize the size of the infected population. Therefore, due to the NPI actions described by the control $u_1(t)$, the susceptible population, illustrated in Fig. 5.2, is still bigger than the noncontrolled path with less infected size, which means

Table 5.1 Values of the scenario simulation

Parameters	Description	Value
β	Transmission rate	5.21×10^{-6}
μ	Recruitment and death rate	1.4167×10^{-5}
λ_1	Quarantined infected individuals rate	0.5
λ_2	Recovered infected individuals	0.42
α_1	Death rate of infected individuals related to the disease	0.1
α_2	Death rate of quarantined individuals related to the disease	0.21
ε	Recovered infected individuals rate	0.5
θ	The rate of contamination	0.04
γ	The clearance rate	0.03
σ	Volatility	0.001

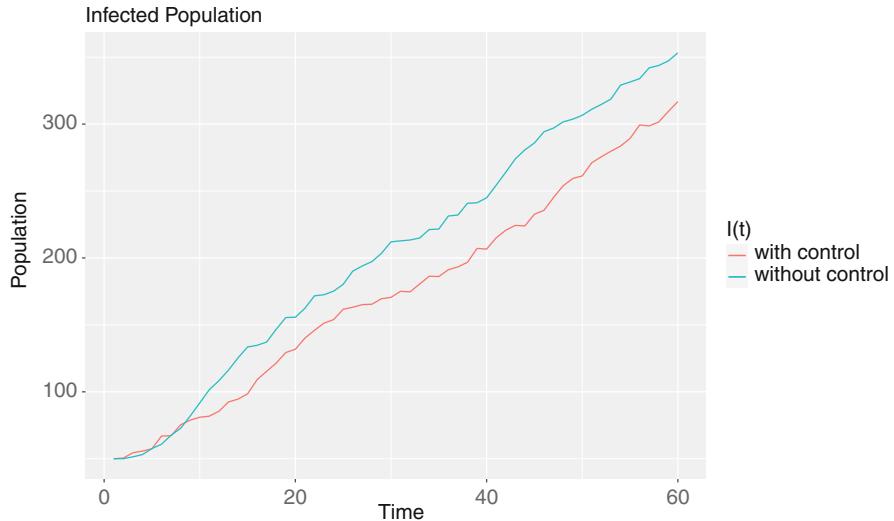


Fig. 5.1 Infected size with and without control strategies

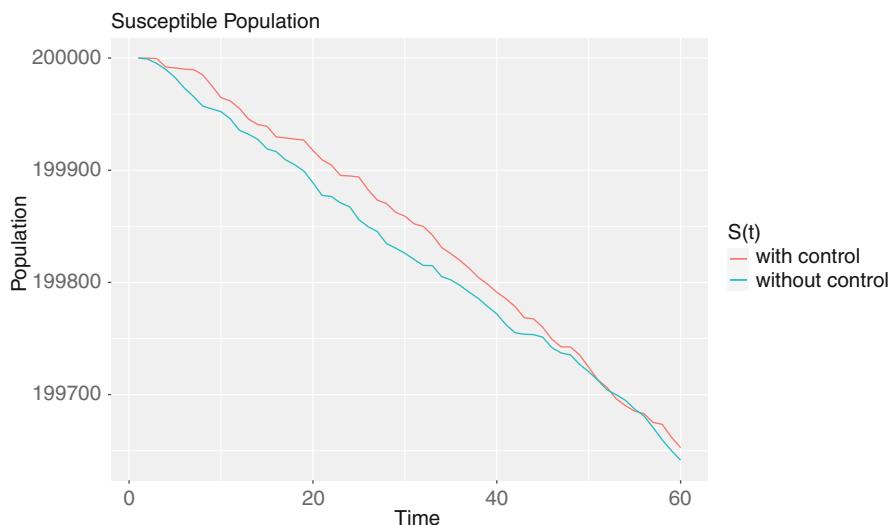


Fig. 5.2 Susceptible size with and without control strategies

less cost for treatment and quarantine. In Fig. 5.3, the quarantine size is lesser with the combination of the measures. Besides, in Fig. 5.4 an augmentation in recovered size will provide reduced stress on the health system, especially in times of healthcare labor shortage. However, the three measures do not impact the

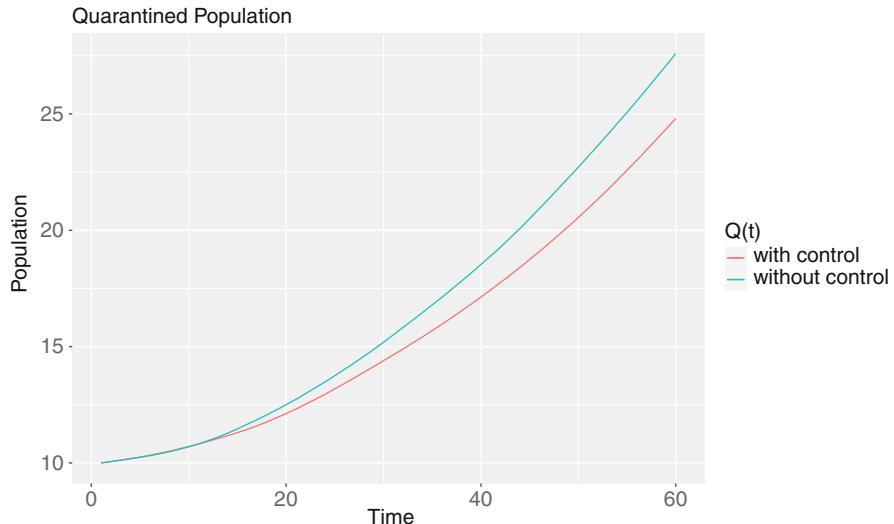


Fig. 5.3 Quarantined size with and without control strategies

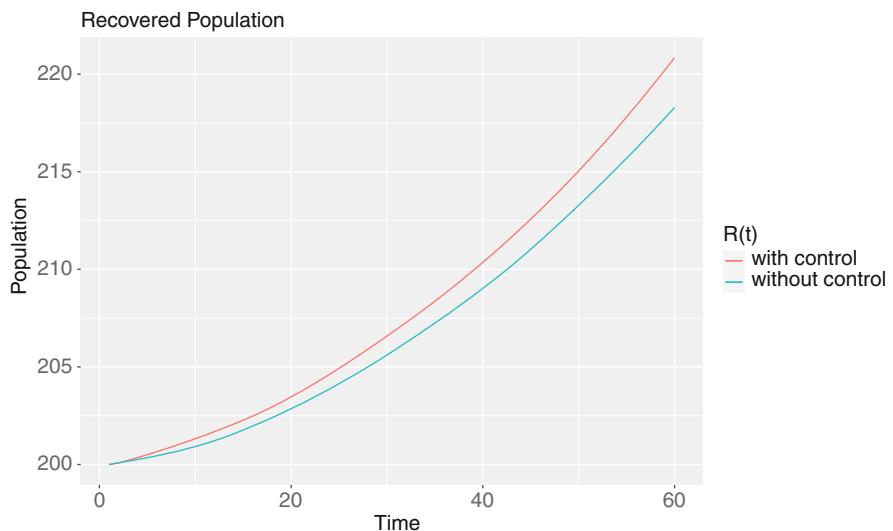


Fig. 5.4 Recovered size with and without control strategies

contaminated environment, as illustrated in Fig. 5.5, which could be related to the chosen value for the parameter of contamination rate θ and an absence of control or strategies on this compartment (Figs. 5.6 and 5.7).

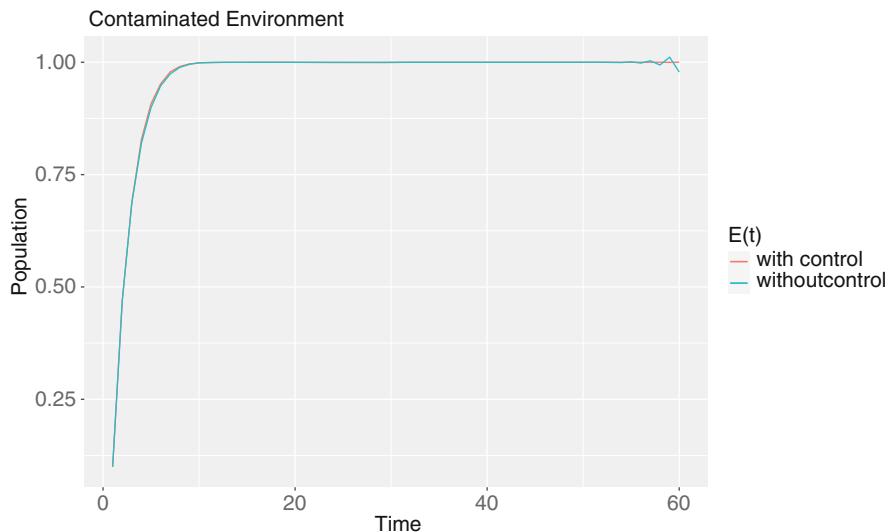


Fig. 5.5 Contaminated environment with and without control strategies

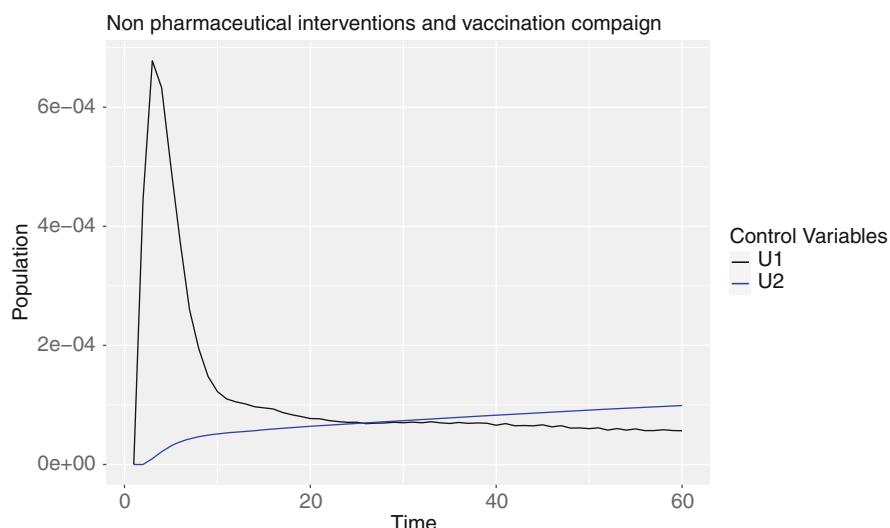


Fig. 5.6 Application of NPI and vaccination strategies

5.4 Conclusion

Viral and epidemic models behave differently and represent a challenge in terms of mathematical analysis and control. Therefore, we seek a combination of these models with an intermediate contaminated environment model, and we include short-

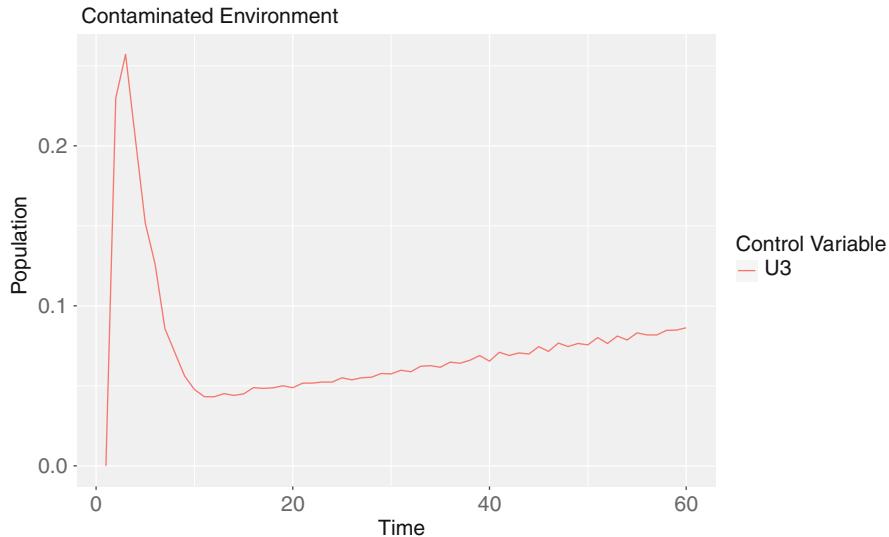


Fig. 5.7 Application of treatment strategy

and long-duration prophylaxis. Motivated by the work of Feng et al. [7], we propose multiple timely controls to describe different measures for both deterministic and stochastic models, where we show the existence of deterministic and stochastic optimal controls for public health interventions. Also, using Pontryagin's maximum principle, we characterize optimal controls for non-pharmaceutical, vaccination campaigns, and treatment strategies. Besides, we conduct in our illustrations a numerical endemic scenario under the condition $\mathcal{R}_0 > 1$. Moreover, we investigate the problem of minimizing the infected size and cost during an epidemic using various measures, both with and without considering the influence of environmental uncertainty. These results reveal the role of combined measures to manage and control an infectious disease. In particular, there is no extensive research on the deterministic and stochastic stability behavior of this kind of coupled model. Therefore, these will be considered in our future investigation.

References

1. Berhazi, B.e., El Fatini, M., Laaribi, A., Pettersson, R.: A stochastic viral infection model driven by lévy noise. *Chaos, Solitons Fractals* **114**, 446–452 (2018)
2. Berhazi, B.e., El Fatini, M., Laaribi, A., Pettersson, R., Taki, R.: A stochastic SIRS epidemic model incorporating media coverage and driven by lévy noise. *Chaos, Solitons Fractals* **105**, 60–68 (2017)
3. Caraballo, T., El Fatini, M., Sekkak, I., Taki, R., Laaribi, A.: A stochastic threshold for an epidemic model with isolation and a non linear incidence. *Commun. Pure Appl. Anal.* **19**(5), 2513 (2020)

4. Van den Driessche, P., Watmough, J.: Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **180**(1-2), 29–48 (2002)
5. El Fatini, M., Pettersson, R., Sekkak, I., Taki, R.: A stochastic analysis for a triple delayed siqr epidemic model with vaccination and elimination strategies. *J. Appl. Math. Comput.* **64**(1), 781–805 (2020)
6. El Fatini, M., Sekkak, I., Laaribi, A., Pettersson, R., Wang, K.: A stochastic threshold of a delayed epidemic model incorporating lévy processes with harmonic mean and vaccination. *Int. J. Biomath.* **13**(07), 2050069 (2020)
7. Feng, Z., Velasco-Hernandez, J., Tapia-Santos, B.: A mathematical model for coupling within-host and between-host dynamics in an environmentally-driven infectious disease. *Math. Biosci.* **241**(1), 49–55 (2013)
8. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A Contain. Papers Math. Phys. Char.* **115**(772), 700–721 (1927)
9. Kirk, D.E.: Optimal Control Theory: An Introduction. Courier Corporation, North Chelmsford (2004)
10. Kushner, H.: Existence results for optimal stochastic controls. *J. Optim. Theory Appl.* **15**(4), 347–359 (1975)
11. Lahrouz, A., Settati, A.: Necessary and sufficient condition for extinction and persistence of SIRS system with random perturbation. *Appl. Math. Comput.* **233**, 10–19 (2014)
12. Liu, Q., Jiang, D.: Threshold behavior in a stochastic SIR epidemic model with logistic birth. *Physica A Stat. Mech. Appl.* **540**, 123488 (2020)
13. Lukes, D.L.: Differential equations: classical to controlled (1982)
14. Nowak, M., May, R.M.: Virus Dynamics: Mathematical Principles of Immunology and Virology: Mathematical Principles of Immunology and Virology. Oxford University Press, UK (2000)
15. Øksendal, B., Sulem, A.: Stochastic Control of Jump Diffusions. Springer, New York (2005)
16. Pitchaimani, M., Devi, M.B.: Stochastic probical strategies in a delay virus infection model to combat covid-19. *Chaos, Solitons Fractals* **152**, 111325 (2021)
17. Rajaji, R., Pitchaimani, M.: Analysis of stochastic viral infection model with immune impairment. *Int. J. Appl. Comput. Math.* **3**(4), 3561–3574 (2017)
18. Rajaji, R., Pitchaimani, M.: Analysis of stochastic viral infection model with lytic and nonlytic immune responses. *Stoch. Anal. Appl.* **38**(3), 490–505 (2020)
19. Wang, K., Jin, Y., Fan, A.: The effect of immune responses in viral infections: A mathematical model view. *Discrete Contin. Syst. B* **19**(10), 3379 (2014)
20. Wodarz, D., Christensen, J.P., Thomsen, A.R.: The importance of lytic and nonlytic immune responses in viral infections. *Trends Immunol.* **23**(4), 194–200 (2002)
21. Yong, J., Zhou, X.Y.: Stochastic Controls: Hamiltonian Systems and HJB Equations, vol. 43. Springer Science & Business Media, Berlin (1999)
22. Zhou, B., Han, B., Jiang, D.: Ergodic property, extinction and density function of a stochastic SIR epidemic model with nonlinear incidence and general stochastic perturbations. *Chaos, Solitons Fractals* **152**, 111338 (2021)

Chapter 6

Modeling Airborne Disease Dynamics: Progress and Questions



Arnab Mukherjee, Saptarshi Basu, Shubham Sharma,
and Swetaprovo Chaudhuri

6.1 Introduction

The most recent coronavirus outbreak (COVID-19) will continue to serve as a stark reminder of the severe impact of highly infectious diseases on society—from the toll on human life to macroeconomics of nations. Health organizations initially believed that the disease spread involved in such outbreaks occurred mostly through direct transmission [75] between individuals, either in the form of physical contact or being close enough for large droplets (ejected during respiratory events) to follow a ballistic trajectory to the susceptible [46, 61]. Indirect transmission through fomites has also been noted [70]. Subsequently, a renewed focus has been put on the aerosol route of transmission, where the small droplets (aerosols) evaporate and remain airborne as desiccated nuclei for long periods of time and have a chance of causing infection once inhaled. It has been argued that the aerosol mode of transmission is the dominant route for the spreading of SARS-CoV-2 [16, 39, 72, 82]. The importance of this route for other airborne diseases such as influenza has also been previously ascertained [22].

In light of these studies and their importance when it comes to the COVID-19 pandemic, we focus on modeling the airborne transmission of viruses in this chapter. The primary question that arises during such crisis is how we can apply the knowledge from the study of disease dynamics to devise a possible approach to “predict” the spread. To achieve this objective, we need to analyze the disease dynamics on a population scale, which is usually done through a statistical

A. Mukherjee (✉) · S. Chaudhuri
Institute for Aerospace Studies, University of Toronto, Toronto, ON, Canada
e-mail: arnab.mukherjee@mail.utoronto.ca

S. Basu · S. Sharma
Indian Institute of Science, Bengaluru, India

description of quantities like the reproduction number R_0 (which is a measure of the average number of secondary infections caused by a single infected individual during the entire course of their infection), or through a more (spatiotemporally) localized quantity Z , which is the number of secondary infections caused by one person during a certain time interval within a certain location. The building blocks of such a statistical analysis would involve modeling the airborne transmission of virus from an infectious person to one or many susceptibles.

A model that encompasses disease dynamics from the very fundamental processes would involve the assimilation of several sub-models each capturing different physical or biological factors that influence disease transmission. We can broadly separate the model into four important constituents (Fig. 6.1). At first, we have the infectious individual and the amount of virus genetic material they carry and shed (through respiratory processes), which is determined through viral load measures and how this load varies over the duration of infection (Sect. 6.2). These viral matters are transported through the respiratory liquid ejected (aerosols) by a person. The capability of the ejected virus-laden aerosols in causing infection is directly dependent on the virus lifetime and behavior of the ejected aerosols under the influence of environmental factors. A primary parameter that dictates the physics (evaporation, drag, gravitational settling, dispersion) during the transmission process is the size of the aerosols. Hence, the second topic of focus is the aerosol size distribution (Sect. 6.3), where we look at some of the significant studies in this field and how the size distribution is modeled based on the observed statistics along with the biological sources of aerosol generation. The logical next step is then to study the airborne transmission process (Sect. 6.4), which is generally split into two parts - the initial transient jet/puff regime, and the subsequent late-time diffusion regime. This airborne duration is generally split into two parts—the transient jet/puff-like behavior which is then succeeded by a late time turbulent diffusion. The different schools of thought in modeling this phase are described along with the various levels of complexities that may or may not be essential to modeling. The final sub-model involves the interaction of the surviving virus-laden aerosols with the susceptible person's physiology—this finally controls the probability of infection, which is described by a dose-response model (Sect. 6.6). Additionally, we take a look at disease transmission through the droplets that have settled on surfaces in

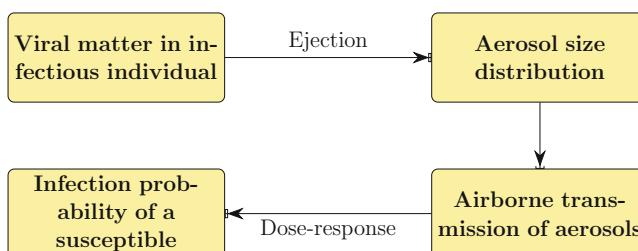


Fig. 6.1 Sub-models describing the airborne disease transmission process

Sect. 6.5 to round up the entire transmission process. In this chapter we discuss the existing literature in context of the COVID-19 pandemic and shed light on the major uncertainties that remain as obstacles in such a modeling approach.

To conclude our discussion, we present (Sect. 6.7) a complete model for the entire infection spread process and use it to generate a distribution for the number of secondary infections. It attempts to capture the underlying physics of the transmission process while introducing the biological aspects through input quantities and a dose-response model. The model has certain simplifications to retain tractability yet maintaining robustness but also has areas for further improvement and study.

6.2 Viral Matter in an Infectious Individual

The disease transmission process starts with an infectious individual, often irrespective of whether they are symptomatic or asymptomatic. The capability of such a person in spreading the disease depends primarily on their viral load, which refers to the amount of virus genetic samples (RNA for SARS-CoV-2) per unit volume of fluid (blood, mucus, etc.) that lie within the individual. The study by Bhavnani et al. [8] showed a direct correlation between higher number of infections developing in contacts and higher viral load individuals, thus indicating that a higher viral load would make one a better agent for disease transmission.

The viral load quantity has been studied in detail for different viruses over the years, but because of the recent pandemic, here we will take a look at the progress made on the viral load description for SARS-CoV-2. The direct correlation between infectivity and viral load for SARS-CoV-2 was more formally shown by Kawasui et al. [51] where they compared the viral load in index and non-index cases (index cases referring to patients who are responsible for at least one secondary infection) and observed their variation with days after infection. It is to be noted that the viral loads considered are specific to the location of biological sample collection, which is often the nasopharyngeal region for SARS-CoV-2. Their results showed higher median viral loads for index cases during sample collection, along with peak values being obtained just after symptom onset followed by a monotonically decreasing tendency over time (in days). Similar observations regarding viral load variation with time were also reported by Challenger et al. [13] where they additionally performed a mechanistic modeling of the process. They divided the entirety of the temporal evolution in two segments—(1) starting with an exponential growth in viral load during the virus incubation phase which eventually reaches a peak due to the resistance caused by the early immune response that attempts to reduce the susceptibility of cells to the virus and (2) followed by the late immune response being eventually effective and alleviating illness by depleting the total mass of infected cells. Several other studies such as those by Kang et al. [50] and Li et al. [56] have also reported (for different SARS-CoV-2 variants) viral load versus time variation to follow a similar sharp increase followed by gradual decrease trend. Figure 6.2 describes this viral load variation, as presented by Jüni et al. [49]. Instead,

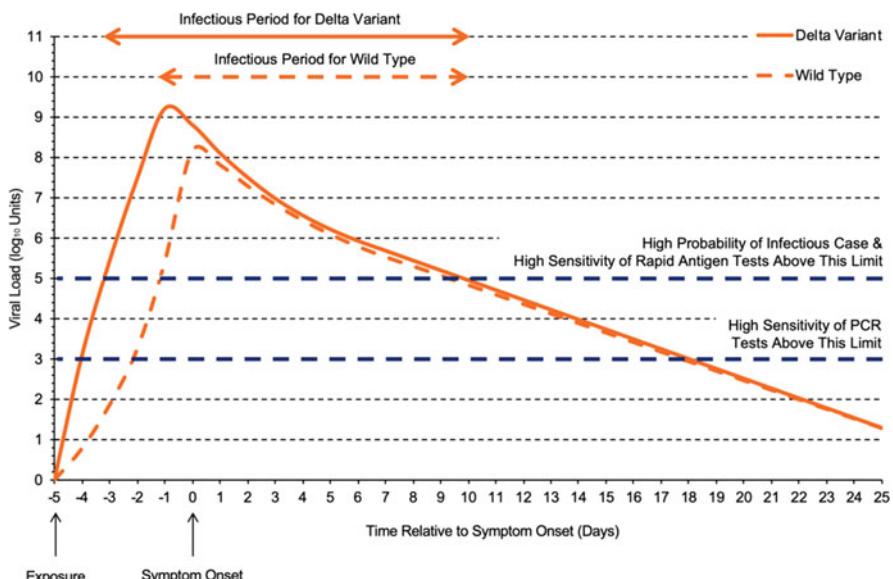


Fig. 6.2 Comparison of SARS-CoV-2 viral load variation (over time) curve for the delta variant compared to the wild type. The graph is generated based on the data of Kang et al. [50] and Li et al. [56]. Figure obtained from Jüni et al. [49] (Figure 1) licensed under Creative Commons Attribution License 4.0 (CC BY)

if a very large number of viral load values are obtained from a very large number of infectious individuals irrespective of the day from the onset of the illness, a time-independent distribution of viral load can be obtained. The same argument goes for removing biases in age (shown by Euser et al. [34]), sex, and biological conditions.

To practically obtain viral load, laboratory techniques like the real-time reverse transcription polymerase chain reaction (rRT-PCR) test (used for COVID-19) are applied on patients, which operates on the principle of identifying the existence of RNA (or DNA) through its amplification using certain biochemical processes [83]. The result of this test is quantified through a parameter termed as the cycle threshold (C_t) value that specifies the number of repeated cycles of amplification of the viral nucleic acid that has taken place (visualized through a fluorescent signal emitted by attached nucleic acid probes) until the signal exceeds a known threshold. This value is inversely proportional to the number of virus copies present because more number of amplification cycles till threshold value would suggest an initially lower viral concentration. Additionally, C_t values are only a relative measure of the viral load—3.3 cycle threshold value difference corresponds to one \log_{10} change

(continued)

in viral load [84]. This is converted to absolute values of viral load through known standard values for the sample being used. The final result is presented in the form of viral genes per unit volume and can be used to calculate the total number of virus particles ejected by an individual.

Therefore, it is also of interest to look at viral load variation within the population as a whole, regardless of the stage of illness an individual is at. For SARS-CoV-2, Yang et al. [110] analyzed a large set of data from testing performed within the University of Colorado campus from presymptomatic and asymptomatic individuals. The viral load values calculated ranged from as low as 8 copies/mL to as high as 6.1×10^{12} copies/mL while fitting a lognormal distribution with a mean of 9.9×10^8 copies/mL (Fig. 6.3). They reported a striking observation that “just 2% of SARS-CoV-2 positive individuals carry 90% of the virus circulating in communities” which presents itself as the long tail of a lognormal distribution. This long-tailed behavior will not be reflected in a time-varying plot of the mean viral load due to the averaging operation over the population. Hence, a relevant focus for future studies could be to observe the viral load distribution (across the population) on each day from symptom onset and compare it to the time-independent viral load distribution. Additionally, the long-tailed nature sheds light on the existence of superspreaders (individuals with extremely high viral load) within the population while suggesting that viral load could possibly be a crucial parameter that governs the number of secondary infections.

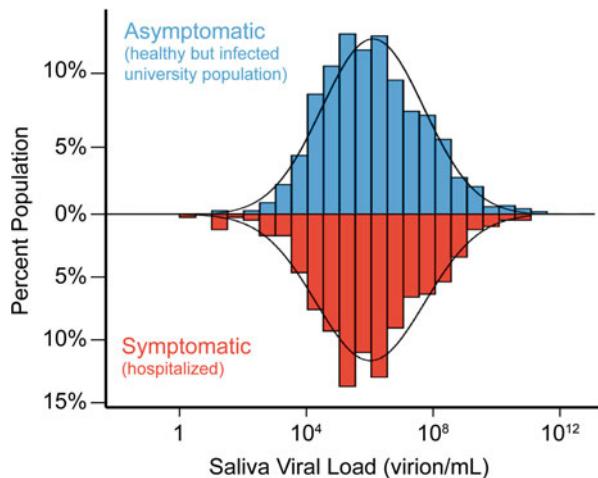


Fig. 6.3 Lognormal density function fitted over a histogram of viral loads in the asymptomatic population compared to the same histogram of viral loads from symptomatic individuals within the University of Colorado campus. Figure obtained from Yang et al. [110] (Figure 2(A)) published and distributed by PNAS licensed under Creative Commons Attribution License 4.0 (CC BY)

6.3 Aerosol Size Distribution in Human Exhalations

The distinction between droplets and aerosols is usually made based on a size threshold. This varies between different studies, but the World Health Organization made the cutoff at a diameter of $5\text{ }\mu\text{m}$ [76] below which the terminology of aerosols is used. In this chapter, we generally use the nomenclature of aerosols for droplets below the $100\text{ }\mu\text{m}$ initial diameter size as recommended by Prather et al. [82] except for scenarios where the size distinction between droplets and aerosols is of direct impact on the topic being discussed, or when we are referring to the entire size range of ejected particles.

The virus particles within the infectious individuals are ejected while being embedded within the generated aerosols during various respiratory events—breathing, talking, coughing, and sneezing. The biological mechanism involved in the case of breathing was recently described by Johnson and Morawska [48] through a bronchiole fluid film burst (BFFB) model. It tells us that during the exhalation process, due to the physiology of breathing, a fluid blockage occurs in the bronchioles. In the subsequent inhalation process, the widening bronchiole will cause the fluid blockage to take the form of a thin membrane or a bubble, which will eventually burst and generate aerosols. This model did not suffer from the inconsistency faced by turbulence-based aerosol generation theories [79] which did not conform with the observation that rapid inhalation leads to a higher aerosol concentration when compared to rapid exhalation. In a subsequent study, Johnson et al. [47] associated the laryngeal region with the aerosolization process during coughing and singing, and the oral cavity with speaking and coughing. Dhand and Li [27] suggested multiple possible mechanisms being responsible for aerosol formation—(1) exiting high-speed air creating instabilities on the mucus-air interface within the larynx, (2) instabilities due to vibration of the vocal folds, and (3) a mechanism analogous to the BFFB model. Recently, the work by Abkarian and Stone [1] shed light on the importance of fragmentation of saliva filaments in the oral cavity in the generation of speech aerosols. The particles generated during these expiratory events encompass a large range of sizes (from submicron range to greater than $100\text{ }\mu\text{m}$), which influences several relevant factors in the transmission physics such as volume of respiratory fluid ejected, evaporation and settling of particles, forces acting on the particles, dispersion, and wall deposition. The information about the range of sizes and their corresponding number and volume contributions are embedded in the particle size distributions (PSD) and volume size distributions (VSD), respectively, and have been studied in detail over the years.

One of the first studies in an attempt to record and analyze particle size distributions was performed by Duguid [30] where the number of droplets ejected during a breath, strong nasal expiration, cough, laughter, and speech was estimated

along with the contribution of the site of origin—nose or throat. In a subsequent study by Duguid [31], the focus was put on the sizes of the droplets ejected along with the corresponding number of droplets. In the calculation of the size of large droplets, they used stains of dye that was left on a slide exposed to the ejected spray. For the smaller droplet nuclei, air sampling was performed using slides (coated with oil) placed in a slit sampler. Using the collection of droplet sizes, a particle size distribution was reported. The primary findings [31] were that the droplet sizes for speaking, coughing, and sneezing ranged from $1\text{ }\mu\text{m}$ to $2000\text{ }\mu\text{m}$. The highest particle counts were reported at the lower end of the spectrum with the range of $2\text{--}100\text{ }\mu\text{m}$ accommodating 95% of the particles, within which the majority resided in the $4\text{--}8\text{ }\mu\text{m}$ range. The average number of particles observed was 1×10^6 for sneeze, 5×10^3 for cough, and 250 for counting loudly from 1 to 100. Similar techniques were applied by Loudon and Roberts [60] in their study involving multiple subjects undergoing speaking and coughing events. They found the average number of droplets across 90 coughing events to be 465, whereas the number jumped to 1764 in the case of loud counting (a difference of at least 1 order of magnitude can be seen in both observations when compared to that of Duguid [31]). Additionally, they were able to ascribe a known distribution (lognormal) to their data. Chao et al. [14] studied speech and cough PSD from 11 volunteers where they applied interferometric Mie imaging technique to enable measurements close to the mouth. They performed numerical simulations to justify that measurements taken at 10mm from the mouth essentially correspond to the original size distribution. They reported that the estimated number of droplets per cough ranged from 947–2085 (geometric mean diameter of $13.5\text{ }\mu\text{m}$), whereas a range of 112–6720 (geometric mean diameter of $16\text{ }\mu\text{m}$) was noted for speech. When compared to Duguid [31], they observed a similar size range containing the highest number of particles ($4\text{--}8\text{ }\mu\text{m}$ compared to $8\text{--}16\text{ }\mu\text{m}$ for Duguid). In an attempt to avoid inconsistencies caused by the introduction of dyes in the mouth during particle size calculations, Xie et al. [108] used glass slides attached to walls along with a microscope and an aerosol spectrometer to obtain PSD from speech and cough. For speech they reported 82% of the droplets to be concentrated below the $100\text{ }\mu\text{m}$ mark, but only 3% were below the $20\text{ }\mu\text{m}$ threshold. For coughs, the corresponding numbers were 64% and 2.5%, respectively. They noted that presence of food dye during the experiment led to the generation of more droplets. For sneezes we can refer to the study by Han et al. [43], where, based on the data from 44 sneezing events, they found two distinct types of size distributions, unimodal and bimodal, and were able to assign a lognormal distribution to each of the modes observed. Their comparison with Duguid [31] showed that they observed 31.2% of the droplets to lie in the $80\text{--}100\text{ }\mu\text{m}$ range as opposed to 34.9% of the droplets lying in the $4\text{--}8\text{ }\mu\text{m}$ range for Duguid. Similar comparisons were also found with the study by Chao et al. [14]. They suggested that the differences observed between their current work and older studies could be a result of possible effects of evaporation (in past studies), lack of collection device used in their present work such that droplets do not impact with anything, and possible differences due to the variability in respiratory tendencies of people along with differences in aerosol generation mechanisms.

In contrast to these studies, Papineni and Rosenthal [78] found all particles for the case of normal breathing to be below the $8\text{ }\mu\text{m}$ limit. Nicas et al. [74] pointed out that this statistic is particularly relevant because the bulk of the volume ejected is concentrated in the size range greater than $8\text{ }\mu\text{m}$ for the cases of Duguid [31] and Loudon and Roberts [60]. Limitations of measuring large droplet sizes with an optical particle counter (as used by Papineni and Rosenthal [78]) have been suggested as a justification for the observed contrast [74]. Focusing primarily on the smaller size range of $0.5\text{--}10\text{ }\mu\text{m}$, Alsved et al. [4] provided PSD for singing, talking, and breathing. The observed median values for the particle emission rates varied from 135 particles/s for breathing up to 1480 particles/s for loud singing. Repeating the expiratory events with face masks appeared to significantly reduce the emission rate. Additionally, a direct proportionality between loudness and emission rate was observed. Asadi et al. [6] used an aerodynamic particle sizer, a spectrometer used for the purpose of capturing particle diameters using their flight duration and the intensity of scattered light, to measure PSD based on different vocalizations and loudness. They confirmed the results of Alsved et al. [4] by showing a linear correlation between loudness of vocalization and particle emission rate (independent of the language spoken). On the other hand, size distribution showed negligible dependence on loudness. This research also shed light on the existence of speech “super-emitters”—people who emit significantly high number of particles compared to others. The physiological factors that dictate this characteristic require further studies. This study was complemented by the findings of Gregson et al. [40] who reported particle size data for singing, speaking, and breathing and found the results to fit a bimodal lognormal distribution. Their experiments were conducted in a zero-background noise setting using an aerodynamic particle sizer, which allowed them to focus on the effects of different vocalizations and loudness. The results agreed with previous studies of Alsved et al. [4] and Asadi et al. [6] and additionally concluded that even though vocalization is a factor to consider when it comes to aerosol generation, the effect of loudness overshadows it.

Most of these studies point toward the PSD having a multimodal shape. As seen before, a unimodal description was fit by Loudon and Roberts [60] to their data. Later Nicas et al. [74] were able to fit a lognormal description to the data of Duguid [31]. Johnson et al. [47] then attempted to formulate this behavior using their BLO model, which is a trimodal formulation where each of the modes (B (bronchiolar), L (laryngeal), and O (oral)) corresponds to three different processes involved in aerosol generation associated with the sites that the name of the model is comprised of. Each mode is described by a lognormal distribution with appropriate constants specific to the mode. Comparison between existing data and the BLO model is shown in Fig. 6.4. This approach was further improved upon by Pöhlker et al. [81] who presented a comprehensive n -mode lognormal formulation where n varies based on the expiratory event in question. They ascribed two modes to breathing (both bronchiolar) with classification between tidal breathing and airway closure breathing. For speaking and singing in addition to the existing bronchiolar modes, one mode corresponding to the larynx and trachea and two oral cavity modes (responsible for generation of large droplets) were assigned. For violent

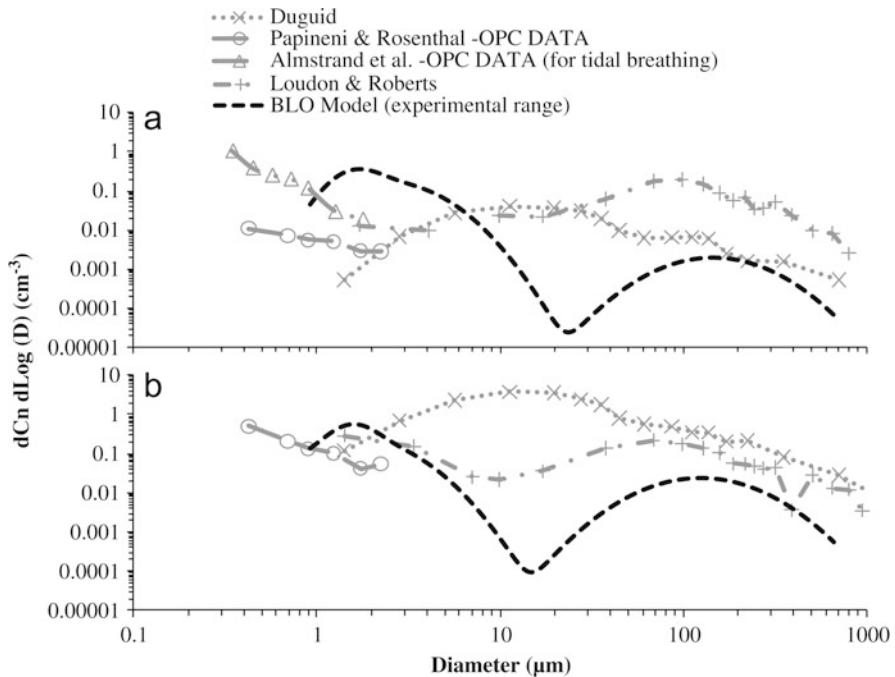


Fig. 6.4 Comparison of size distributions (C_n is the number concentration and D is the particle diameter) based on the BLO model (Johnson et al. [47]) with the results reported by Duguid [31], Loudon and Roberts [60], Papineni and Rosenthal [78], and Almstrand et al. [2] (only tidal breathing) for (a) speaking and (b) coughing. Reprinted from Johnson et al. [47] with permission from Elsevier

expiratory events like coughing and singing, the same five modes were applied. The formulation is given as

$$\frac{d\eta}{dD_{s,0}} = \log_{10}(e) \sum_{i=1}^n \frac{A_i}{D_{s,0}} \exp \left(- \left[\frac{\ln(D_{s,0}/D_i)}{\sigma_i} \right]^2 \right) \quad (6.1)$$

where η describes the number concentration, $D_{s,0}$ is the sample space variable corresponding to the particle diameter at ejection, and A_i and σ_i are event-specific constants. This description is used in the model presented at the end of this chapter.

Certain additional considerations need to be made when calculating or using size distributions. First is the effect of evaporation water loss on the ejected particles. The size distributions calculated are often at a certain distance from the mouth and can be expected to underestimate the original PSD at the source. Studies have often applied corrections to the observed distribution (Duguid [31] applied a fourfold correction factor) or used numerical methods to justify their observations [14]. Nicas et al. [74] provided a simple model for obtaining equilibrium sizes after complete desiccation

of the droplets. They equated the mass of nonvolatile substances within the droplets prior to and post-evaporation to obtain a 0.44 times reduction in diameter. In the case of incomplete desiccation, they calculated a range of factors based on the relative humidity and landed on 0.5 being the most appropriate value given the uncertainties associated with it. The evaporation effect can also be tracked over time to account for the modified droplet size as shown by Chaudhuri et al. [15, 16], where they coupled evaporation physics along with a model for the subsequent crystallization process to calculate the instantaneous particle size. Another question that specifically appears when dealing with infectious aerosols relates to whether the distribution for aerosols with virus particles is same as the original aerosol PSD. This particular topic was looked at in the recent work by Anand and Mayya [5], where their analysis suggested that the virusol (aerosol with virus) distribution retained the lognormal nature of PSD but the mean shifted toward higher diameters when compared to the original distribution. A particular observation of note was that for mild cases (viral load below 10^5 mL/copies), only 0.1% of the particles with diameters below 60 μm contained virus particles, whereas larger droplets would undergo gravitational settling in a rapid manner. This suggests that mild cases are often incapable of spreading infection through the aerosol transmission route and instead points toward high viral load individuals behaving as “superspreaders.” But a recent study by Coleman et al. [21] measured virus RNA in aerosols ejected by (SARS-CoV-2) infected individuals, where their observations contradicted the analysis by Anand and Mayya [5] and showed that 85% of the ejected viral matter resided in the finer aerosol particles (diameter less than 5 μm). This difference emphasizes the uncertainty when it comes to understanding the contribution of different particle sizes in carrying viral matter and suggests that more detailed studies on this aspect is required for more conclusive theories.

6.4 Airborne Transmission of Aerosols

The infectious aerosols ejected from an individual will only be capable of inducing illness after it reaches the target susceptible. The intermediate steps in transmission are governed by complex multiphase fluid dynamics and its description is subject to differing approaches. If our aim was only to accurately determine the number of people that would get infected within a region of interest under given conditions, without any consideration of computational cost, it would suffice to use the most exhaustive methods to model this process of transmission, i.e., direct numerical simulations (DNS). In DNS applications all flow scales are resolved, which, especially when turbulence is involved, become extremely expensive to use even for a singular case. This makes it untenable for problems that require the analysis of a large number of such cases, which is expected to happen when the intention is to study pandemic dynamics where our desired result is a distribution of secondary infections. The central question that arises here is how much complexity is required to make an impact on a statistical analysis of infection spread, as conserving

resources required to solve such problems is always a consideration. Based on this we can have a hierarchy of complexity when it comes to modeling the transmission of ejected aerosols. At the highest level would be a fully inhomogeneous exhaled jet with local clustering of aerosols that travels through the air under the influence of drag, gravitational, and buoyant forces and is subjected to evaporative effects. This requires rigorous numerical simulations to solve. After one set of simplifications, we can consider a jet of aerosols that travels under similar conditions as before, but the aerosol particles that make up the puff/cloud are assumed to be homogeneously distributed, thus ensuring that any person that comes in contact with the jet has equal chances of infection (biological differences notwithstanding). This stage is usually dealt with multiple equation models that track the changes of various parameters that describe the puff. Further assumptions can allow us to neglect the jet flow entirely due to its transient nature and attribute the primary force behind the infection to the turbulent diffusion of the aerosol cloud. A one equation model often suffices to represent this stage. Finally, at the simplest level, a homogeneous distribution of infectious aerosols can be considered. All four levels of complexity will be discussed in more detail in this section.

For a virus-laden particle to infect, it has to either directly make contact with a nearby susceptible before settling or stay airborne long enough for it to be inhaled. This distinction was first made in the study by Wells [105] in 1934, where the stark difference in transmission dynamics of large and small droplets was shown—droplets above the $100\text{ }\mu\text{m}$ range are prone to rapid gravitational settling (within seconds) as opposed to those below this limit (aerosols) which tend to evaporate and remain airborne as desiccated droplet nuclei for large periods of time. This classification splits disease transmission into two major routes, direct droplet transmission (for large droplets) and indirect aerosol transmission due to suspended desiccated droplets in air, and has been the method of tackling aerosol transmission for years. The main points of contention with the dual routes of transmission approach were summarized by Bourouiba [10]. The first of which is the use of infection data from the population (as opposed to doing so from studies specifically focused on host-to-host transmission) by the health sector as a way to deduce which mode of transmission is active for a certain infectious disease. The other issue is that droplets are not individually emitted such that they can either undergo rapid settling or remain airborne as aerosols. They are ejected as a part of a turbulent multiphase puff that imparts its own characteristics on the settling and evaporation of the droplets. This could be through carrying droplets to further distances due to the inherent momentum of the puff or through delaying evaporation due to differing conditions within the turbulent puff when compared to the ambient. However, some of the largest droplets exit the puff right after ejection without being influenced by the puff dynamics.

An exhaustive model for such turbulent puffs was provided by Bourouiba et al. [11] where they investigated the aerodynamics of expiratory events like coughs and sneezes. They described these flows as turbulent, multiphase, buoyant flows and presented a six-equation model, the results of which strongly matched with their experimental findings. The model involves a homogeneous puff/cloud assumption

where the cloud density evolves with time, due to entrainment, as a function of different ratios of fluid and particle density. The trajectory of the turbulent puff was determined by tracking the changes in its initial momentum and buoyancy. Based on the foundation laid by Martin and Nokes [68], the final complexity introduced by Bourouiba et al. [11] was that of particle fallout with time, under the assumption that the cloud is well-mixed and fallout occurs only through the bottom half of the cloud. In the same study, Bourouiba et al. [11] experimentally showed that the dynamics of this turbulent puff can be broadly split into two regimes—the first resembling jet-like behavior for a short duration (i.e., the distance travelled by the cloud evolves as one-half power of time) and the remainder being governed by self-similar growth of the puff [73] (i.e., the distance travelled by the cloud evolves as one-fourth power of time). Based on these observations, two separate entrainment coefficients were calculated, the average of which acted as input to the analytical model in an attempt to capture the effects of both regimes.

For nonhomogeneous treatment of particles within the cloud, we need to turn our attention to more computationally intensive models. Dbouk and Drikakis [23] applied a Reynolds averaged Navier-Stokes equation model for the bulk multiphase flow coupled with a Lagrangian model that tracked the changes in saliva particle characteristics like mass, velocity, temperature, and spatial location. Their model also accounted for interactions between the droplets themselves and with the surrounding fluid. For the case of no wind speed, they found the range of most droplets to be within the 1m distance from ejection point. But for wind speeds between 4km/h and 15km/h, this range increased to 6m. The results of Dbouk and Drikakis [23] were complemented by the findings of Feng et al. [36] and Li et al. [57] where they also demonstrated strong correlation between environmental conditions and distance travelled by droplets. Feng et al. [36] observed that the recirculation zones created in the wakes of the infector and the susceptible often trap the droplets, preventing them from escaping and altering their trajectories drastically. Direct numerical simulations were performed by Fabregat et al. [35] for coughs with the intent to capture the turbulent effects of the puff. Accounting for variations in ejection velocity (initial acceleration followed by deceleration for the remainder of the process), they observed buoyant forces influencing the trajectory of the puff once the deceleration process starts. After the ejection process terminated, the turbulent kinetic energy of the puff rapidly dropped leading to quick equilibration between the puff velocity and the ambient velocities, which in turn opens up the puff to be easily influenced by varying indoor conditions (ventilation, gradient in room temperature, etc.). They also focused on the evolution of the puff topology and determined that an ellipsoidal assumption as opposed to a spherical one is more appropriate. Numerical simulations by Chong et al. [20] found small particles remaining in the cloud for a longer duration due to protection from the surrounding humid puff. Their results also showed that even droplets of initial diameter of $10\text{ }\mu\text{m}$ go through the aerosol transmission route, bringing into question the use of $5\text{ }\mu\text{m}$ or similar values as the cutoff for the droplet-aerosol distinction as used by the World Health Organization [76]. Simulations by Rosti et al. [89] attempted to pinpoint the effects of turbulence on puff and droplet dynamics through

the comparison of high-resolution DNS with corresponding coarse-grained variants that are not able to fully resolve the smallest scales. The primary effect was seen through a delay of evaporation due to turbulence, which in the coarser-grained variants showed up as lighter droplets being transported to larger distances. While this result was found for coughs, it was noted that for higher-density expiration events like sneezes, the influence of droplet clustering could have an effect on the evaporation characteristics. Overall these studies serve to emphasize the importance of coupling the droplet phase motion with the turbulent puff dynamics so as to capture the detailed physics involved and in turn not underestimate the droplet's travel distance.

Moving on from our focus on the transitory jet/puff regime, we turn our attention to the diffusion-dominated regimes that succeeds it. Focusing on only the puff part of the model would inherently assume that the instantaneous puff motion is the primary cause of infections. After some time the cloud's motion would have significantly slowed down and would be comparable to the velocity fluctuations in its surroundings, hence making the turbulent diffusion of aerosols (based on ambient turbulence) the dominant mode of transmission. Balachander et al. [7] provided a highly simplified approach toward calculating the approximate time of transition to the diffusion-dominated regime. Assuming the existence of a threshold time for transition, they equated the cloud radius based on the puff model with that of the cloud radius modeled using Richardson's scaling law [87] (for the diffusion-dominated regime) along with equating the growth rate of the cloud based on the puff model and the diffusion model. Noting that Richardson's scaling law is an overestimation of the diffusive effect, they found that for typical values the transition time is of $O(1)$. For nonviolent expiratory events like speaking and breathing, the initial momentum might be low enough that the puff-like behavior is highly transient and could justify modeling the entire process as being solely diffusive.

The study by Keil [52] demonstrates a hierarchical methodology to modeling aerosol dispersion within a room. The simplest method involves assuming well-mixed conditions within the room, under the assumption that the mixing process is instantaneous. Application of such models involves minimal difficulty, but it comes with the drawback of under-predicting concentrations in certain high-density pockets that may occur within the region of interest (due to interaction of the aerosols with objects or individuals). An improvement to this approach is done through the introduction of spatial variations in concentration. Multiple zone models divide the region of interest into several zones which are individually assumed to be well-mixed, while an exchange parameter is introduced to tackle the mass transfer between zones. Among other uses of this method, Özkaynak et al. [77] used this approach to predict pollutant concentrations in indoor locations. Ryan et al. [91] introduced a sequential box model where each box has its own sources and sinks and interacts with adjacent boxes through transfer of air based on the pattern of air transfer that had been established prior. This reduced the drastic increase in complexity that came with increase in the number of boxes as previously all boxes were made to interact with each other. The sequential box model can be represented as a set of coupled differential equations which under equilibrium assumptions

reduce to a tri-diagonal matrix representation. Finally, one can opt for a diffusion-based model (molecular or turbulent) where the solution to the diffusion equation provides a readily usable analytical formulation. The intrinsic assumption here is that the ambient air does not have strong mean velocities, even though that condition can also be modeled but will only provide situation-specific solutions. Drivas et al. [29] provided an analytical model based on the diffusion equation while accounting for wall reflection terms. The model looked at a point source placed within a room with isotropic turbulence along with ventilation and wall deposition effects being accounted for. The entirety of turbulence is modeled through the introduction of a turbulent diffusivity term. Such treatment of turbulence is applicable with relative confidence when the turbulent length scales are much smaller than the length scales of the region of interest. The solution to the problem is presented as an analytical formula allowing for ease of use. Cheng et al. [19] performed experiments on CO concentrations within an indoor location and fit Drivas et al.'s [29] model to their data to obtain an empirical relation for the turbulent diffusion coefficient. This was shown to be proportional to the square of the room length scale and the air change rate. Venkatram and Weil [103] have reanalyzed gradient transport-based turbulence models and found that it is reasonable to apply them within indoor environments.

All the above approaches toward modeling airborne transmission of droplets and aerosols need to be put in the context of a statistical analysis of secondary infections. We have seen that for the description of turbulent puffs, a homogeneous assumption is usually reasonable to determine the expected range of travel for droplets in the case of no ambient air velocity. DNS approaches do bring in more accurate results but are likely more useful as guidelines toward simpler modeling approaches by unraveling the underlying physical phenomenon. When our scope is over a large number of cases (required for statistical analysis), it can be expected that the average population flow within indoor locations tends more toward individuals spending long enough to be impacted by the diffusive spread of infectious clouds that would dominate shortly after ejection. Therefore, for such cases additional levels of simplifications bring us to purely diffusive models which have been shown to work well. For non-impulsive expiratory events, the transitory jet regime might be short enough to justify such an approach, but more studies are required when it comes to violent expiratory events. Neglecting convective effects is a point of contention but was addressed to an extent in Chaudhuri et al. [17] where they compared the Drivas-Cheng diffusion model [19, 29] with the experimental and computational results of Hathway et al. [44] for CO emissions within a room. The comparisons showed qualitative and quantitative match for most sampling points and make a case for the use of a purely diffusive model. In the same work, Chaudhuri et al. [17] also applied a well-mixed model and found that the probability density function (pdf) generated for secondary infections match with that from a Monte Carlo simulation of the Drivas-Cheng model. These results give the impression that the simpler approaches are often sufficient when one intends to involve statistical techniques because small deviations would be averaged out at the end. Some uncertainties associated with this methodology appear from the interaction of the human body with the infectious aerosols and have been brought up later in this chapter.

6.5 Transmission Through Fomites

Prior to discussing the effect of the airborne virus-laden aerosols on the susceptible, it is prudent to discuss the transmission process through fomites, which are a side effect of large droplets following a ballistic trajectory and depositing on nearby surfaces due to higher inertia and gravitational force. These droplets potentially carry the embedded pathogens (along with other constituents), and subsequently contaminate daily-use surfaces which play a role in infection spread, and are usually termed as fomites. The ingestion of these deposits by a healthy individual through direct or indirect means results in fomites-based disease transmission, depending on the viability of the embedded pathogens. The activity of deposited pathogens is not limited to its chemical, biological, and virological aspects; it also depends on the involved hydrodynamics, interfacial dynamics, and mutual interactions between the pathogens and surrounding mediums during the desiccation of deposited droplets. Here we will only discuss the hydrodynamics and interfacial aspects and how they influence the pathogen distribution, viability, and virulence due to desiccation.

Pattern formation due to the evaporation of sessile droplet dispersed with solid particles is well-studied in the literature. Deegan et al. [24] explained the famous “coffee ring effect” through the capillary flow from the droplet apex to the contact line, forming a concentrated ring deposition of solid particles near the three-phase droplet contact line. A single-ring deposition is observed when the contact line is pinned to the substrate; however, other patterns, such as concentric rings, can also occur due to pinning and de-pinning of the contact line during the evaporation process [63]. The pattern formation is also altered by surface type, dispersed particle shape, surface roughness, and solute-solvent-substrate interactions [18, 65, 90, 111]. The physical mechanism of pattern formation mentioned for the above simple fluids can be translated to respiratory fluids. However, due to their complex composition (such as dispersion of salts, protein, surfactants, pathogens, etc. [102]) and mutual interaction between the constituents and surrounding medium, respiratory fluid requires critical attention. The pathogen’s size mainly varies from the nanometer to micrometer range. Thus they have lower Stokes number values. Hence their deposition is primarily governed by the droplet’s internal flow and pathogen’s motility. The presence of several constituents inside the droplet can alter the relative dominance of capillary flow (driven by continuity) and Marangoni flow (driven by surface tension gradient) and hence the final deposition [86]. Aspects like suppression of coffee ring due to protein adsorption on the dispersed polystyrene particles have been reported [26]. The central or edge deposition was found to be dictated by the protein charge [94]. The presence of salts results in crystal formation on desiccation. Generally, multi-fractal dendritic crystals are formed during biofluid evaporation whose structure depends on the salt concentration, particle morphology, and drying modes [37, 38, 80]. The crystal formation was also shown to be substrate dependent, where regular cruciform or dendritic structures were observed for glass, plastic, and PET substrates and regular cuboidal crystals were formed on steel substrate (see Fig. 6.5, [86]). Irrespective of the surface

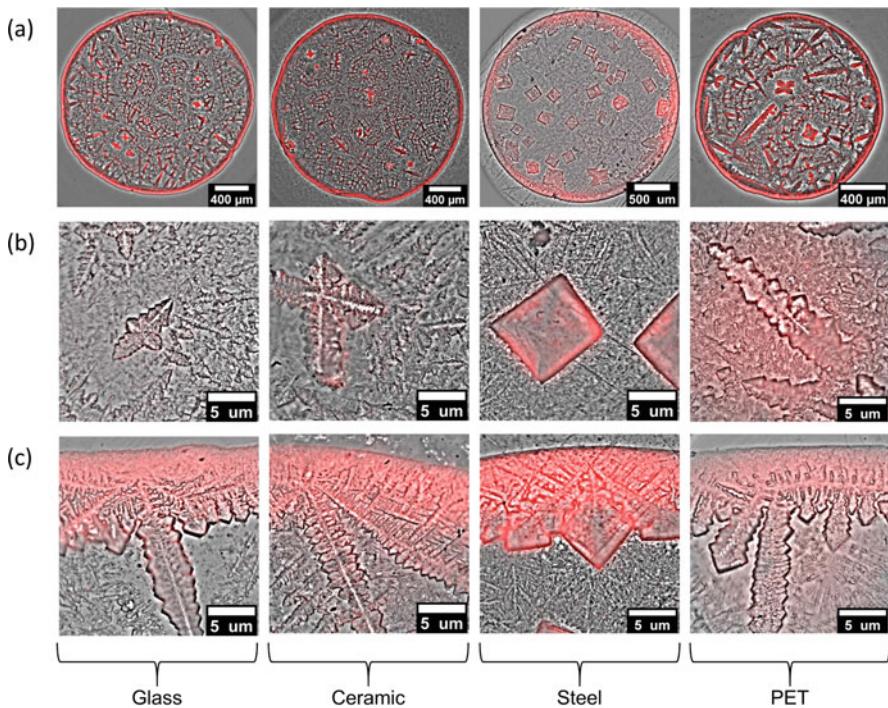


Fig. 6.5 Superimposed fluorescent and bright-field images of the droplet deposition pattern on the glass, ceramic, steel, and PET substrate from left to right. **(a)** Zoomed-out view of deposition. **(b)** Central region. **(c)** Peripheral region. The fluorescent intensity depicts the extent of particle deposition at the end of crystallization. Reprinted from Rasheed et al. [86] with permission from Elsevier

choice, thick edge deposits were observed on all substrates due to dominant capillary flow during most of evaporation period. The formation of dendritic and cubical crystal structures was explained based on the extent of mucin adsorption on different substrates. To identify the locations of pathogen deposition by only considering the hydrodynamic effects, 100 nm fluorescent polystyrene particles were used as virus-like particles (VLPs). Such substitute of actual pathogen for studying hydrodynamic effects has been done on several occasions in literature [62, 95]. The fluorescent signal from VLPs is superimposed on the microscopic images of the desiccated patterns (see Fig. 6.5). The VLPs are primarily deposited near the droplet's periphery or on the crystal edge. The peripheral deposition was due to the capillary flow during evaporation and crystallization stages, while deposition near crystal edges was attributed to the jet flow toward crystal which arises during the crystallization stage [32]. Using scanning electron microscopy (SEM) images, it was shown that all the dispersed VLPs were infused inside the mucin layer, and no exposure of VLPs on the desiccated pattern was found. Some spherical protrusion of size similar to VLPs was observed; however no clear exposed

VLPs could be identified. A recipe for preparing surrogate respiratory fluid using an aqueous solution of NaCl (salt), mucin, and 1,2-dihexadecanoyl-sn-glycero-3-phosphocholine (DPPC) was provided by Vejerano and Marr [102] and was used for investigating the desiccation of surrogate respiratory droplets under variable humidity conditions. The efflorescence limit at which the nucleation of salt crystals starts was found to be dependent on relative humidity, which further affects the crystallization dynamics. A fluorescent tag lipid (NBD-PC) was used to track the $\phi 6$ virus location inside the droplet. The virus was located uniformly around the droplet (see Fig. 6.6), indicating no specific preference for virus deposition. The crystallized salt and embedded viruses were encapsulated by other constituents (mucin and DPPC), which respond differently upon rehydration. For droplets that don't contain a surfactant, mucin separates from the solution and deposits on the liquid-air interface, forming an encapsulated shell. Such shells get transformed into gels on rehydration. However, for the droplets containing surfactants, it was suggested that some fraction of surfactant gets diffused toward the liquid-air interface and alters the physicochemical characteristics, thereby restricting the rehydration of the deposition upon exposure to saturated humidities. These results suggest restricted rehydration of fomites in case of inhalation, i.e., rehydration due to saturated conditions inside respiratory tracts.

The bacteria or other pathogens present inside the biofluid droplets undergo several stresses, such as shear stress due to the droplet's internal flow, starvation stress due to lack of nutrients, desiccation stress due to depletion of water from bacterial cells, and stress due to environmental conditions (temperature and humidity) during desiccation process [64]. These stresses collectively decrease the viability of the bacteria [109]. However, few bacteria could remain alive for weeks under such conditions on inanimate dry substrates [55]. The bacteria present in pure fluid droplets also result in pattern formation such as "coffee rings" [3, 109] which could be altered by the wettability and morphology of the bacteria [28, 66, 101]. More recently, Rasheed et al. [85] investigated the self-assembly mechanism and morpho-topological changes which arise in *Klebsiella pneumoniae* bacteria suspended in drying sessile droplets. Localized measurements were taken at the central and peripheral regions of the desiccating droplet. More compact assemblies of bacteria with significant bacterial deformation were formed near the droplet periphery, while a cellular pattern with more spatially placed bacteria was observed in the central region. This morpho-topological arrangement of bacteria has a direct implication on the viability of bacteria. The bacteria near the droplet periphery are less exposed to desiccation due to their compact assembly and hence are more viable when compared to the bacteria deposited in central regions. The results from this section indicate the influence of hydrodynamics and interfacial dynamics on the pattern formation and survivability of the pathogens during the desiccation of respiratory droplets, which has a direct implication on the efficiency of infectious disease transmission through fomites.

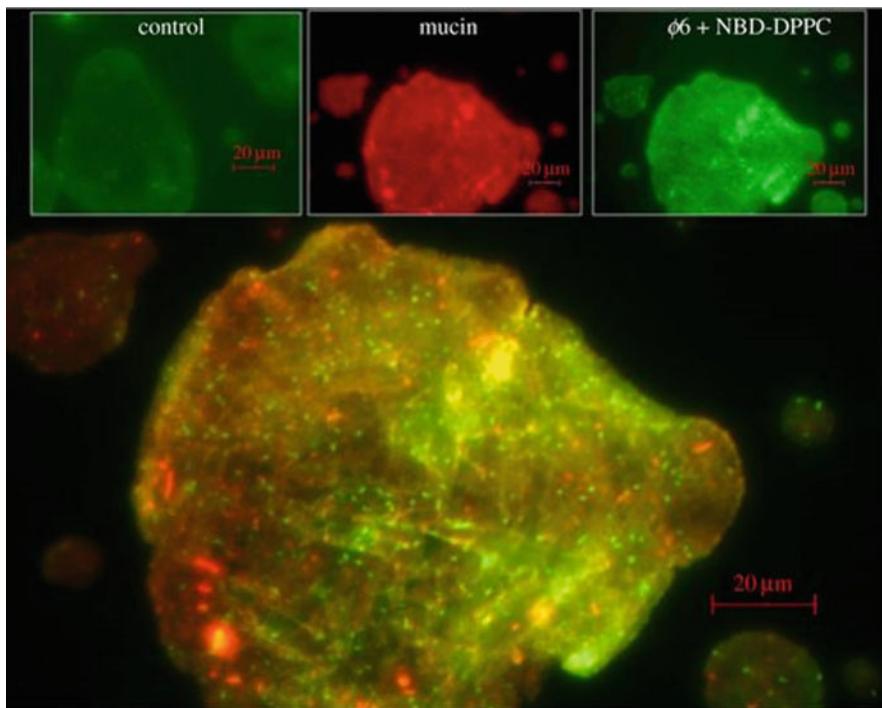


Fig. 6.6 Composite fluorescent image of a 4C droplet containing $\phi 6$ virus exposed to 29% RH. The bright green dots approximately 1 μm in size may indicate the location of the virus. All scale bars are 20 μm . Reprinted from Vejerano and Marr [102] with permission from RSIF licensed under [Creative Commons Attribution License 4.0 \(CC BY\)](#)

6.6 Infection Probability of a Susceptible

Till now we have had the virus come in contact with the susceptible either through existing as embedded matter within the ejected mucosalivary liquid that has travelled the distance between the index case and the susceptible or through fomites. But the act of infectious aerosols reaching the target susceptible is only part of the transmission dynamics that would finally determine whether infection will be induced or not. Here three factors must be considered—(1) the amount of virus particles that have survived within the infectious aerosols, (2) the number of surviving virus particles that are inhaled, and (3) the human body's response to the inhaled virus and the subsequent possibility of illness.

Whether virus particles embedded in aerosols will survive the journey from the ejection source to the susceptible individual depends on their behavior in the presence of varying environmental conditions. Studying the effect of temperature on SARS during the 2003 epidemic in Beijing and Hong Kong, Bi et al. [9] observed number of daily cases dropping with an increase in maximum and minimum

temperatures. Tan et al. [99] found a strong correlation between ambient temperature and the number of reported cases (of SARS) while noting that the temperature range of 16°C to 28°C provided the best chances for virus spread. Similar to SARS, for SARS-CoV-2 several studies have also reported an inverse relationship between temperature and number of cases [33, 97]. Some studies have however contradicted these observations [12, 100], such as that of To et al. [100], where their analysis of COVID-19 spread in Canada found a lack of statistically significant correlation between the total number of cases and ambient temperature, suggesting that our understanding of the influence of temperature on virus survivability (for SARS-CoV-2) still retains uncertainties. A similar inverse relation has been observed between humidity and virus survivability. Wu et al. [107] analyzed COVID-19 cases over 166 countries and reported a 0.85% drop in daily new cases corresponding to a 1% rise in relative humidity. This inverse dependency is also noticed by Lin et al. [58] in their study focused on COVID-19 cases in the USA. Marr et al.'s [67] analysis (on influenza) suggested that relative humidity determines a virus particle's stability by dictating the evaporation physics of an aerosol and thus determines its size and composition (and subsequently its internal chemistry). The same study associates the influence of temperature as an important factor governing the chemical kinetics within an aerosol. These effects can be incorporated into a virus survivability function as shown in Chaudhuri et al. [16] with the virus half-life value calculated based on the ambient conditions (using a calculator provided by DHS [45]).

At this stage the infectious aerosols with the surviving virus particles are inhaled by the susceptible. This process can be modeled using known inhalation rates of a person and the duration of the inhalation event. The impact of the virus particles on the body after inhalation and the subsequent probability of causing illness are described by the dose-response (reaction to a stimulus) of the human body. Quantitative risk assessment for airborne diseases is practically performed using either the Wells-Riley model or models with similar approaches. The Wells-Riley model [88, 106] describes an exponential relation between the probability of infection and the generation rate of a quantum of infection (number of infectious airborne particles that are needed to bring about infection in a person [98]) based on Poisson distributed infectious airborne particles along with assumptions of steady-state concentration field in a well-mixed room [98]. Though there have been modifications of the Wells-Riley model over time to remove such assumptions, the more general dose-response models are additionally able to account for factors such as the route of transmission and the associated physics within, along with having the advantage of not relying on a hypothetical unit for infectious dose (quantum). One of the first studies on dose-response based estimation of risk was performed by Haas [41] for waterborne virus and bacteria. The study compared three dose-response models, starting with the lognormal description which functions under the assumption that any host organism possesses a minimum infection dose (assumed to be lognormally distributed within the population), which when exceeded will cause infection. The second model considered was an exponential single-hit model which is based on the probability that a single virus or more reach the target organ

after the susceptible is exposed to the initial dose, so that it can trigger illness. An improvement to this approach constituted the final model, where the infectivity of virus (or, conversely, the host sensitivity) was assumed to possess a beta distribution [71]. The primary observations from the comparison of the models found that both the beta-Poisson model and the lognormal distribution fit a large number of data sets. This dose-response approach was applied for the case of SARS-CoV (in mice) by Watanabe et al. [104] where they fit data to the beta-Poisson and the single-hit exponential model. Both models provided good fits to the data and it was observed that the beta-Poisson model did not provide a statistically significant advantage over the much simpler exponential one. Based on this, Haas [42] provided a framework for modeling the dose-response of SARS-CoV-2 using the exponential model (developed with average dose being represented in plaque-forming units). Schijven et al. [92] modified this to represent average dose in terms of RNA copies of the virus in the form

$$\mathcal{P} = 1 - e^{-r_v \mathcal{N}_v} \quad (6.2)$$

Here \mathcal{P} is the probability of infection, \mathcal{N}_v is the total number of virion copies inhaled, and r_v is the dose-response constant (inverse of the probability that a single virus will survive long enough to induce illness [42]). The probability of infection \mathcal{P} has two different interpretations based on the description of \mathcal{N}_v . If we express the infectious dose as a function of space and time $\mathcal{N}_v(x, y, z, t)$, then \mathcal{P} dictates the probability of a single individual at the spatial location (x, y, z) at time t being infected. Instead, if we use an average infectious dose $\mathcal{N}_v(\tau)$ that any susceptible person is exposed to at a location (where τ is the exposure duration), \mathcal{P} can be thought of as the average proportion of susceptibles that will become infected once exposed to the average initial dose [42].

Here, \mathcal{N}_v incorporates all the physics-based aspects discussed in the previous sections, whereas the dose-response constant integrates the biological factors. Even though Schijven et al. [92] used a constant value of the dose-response constant r_v for their analysis, there exists a lot of uncertainty in the exact value for this quantity (for SARS-CoV-2) in existing literature. The dose-response constant as defined by Haas [42] was based on viable infectious virus and thus can be found through the dose-response model if infection data (based on dosage of viable virus) is available. Connecting this quantity to the virus RNA copies present in the aerosols being inhaled requires an additional factor that provides the ratio between viable infectious virus (in plaque-forming units or pfu) and the number of RNA copies—also known as the infectivity of the virus.

Plaque-forming unit (pfu) is a commonly used measure for infectivity that represents the number of infected cell regions or plaques formed by a virus. It is analogous to units such as TCID₅₀ which stands for tissue culture infectious

(continued)

dose that will lead to 50% of a cell culture being infected. The primary difference of note is that pfu is a quantitative measure whereas TCID is more qualitative in nature (as it measures the ratio between infected and uninfected cells) [96]. In practice, a constant factor is often found that varies from virus to virus and can be used to bridge the gap between these two measures.

Schijven et al.'s [92] study found 1 out of 80 RNA copies capable of infecting, providing one of the highest infectivity values reported for SARS-CoV-2. Sender et al. [93] write that infectious virion copies ranging from 10^5 to 10^7 were found corresponding to a total virion count of 10^9 to 10^{11} . Despres et al. [25] reported infectivity values for the Alpha, Epsilon, and Delta variants of SARS-CoV-2 and found that the delta variant was on average most infectious while the alpha variant came last. Their observed values for infectivity ratios varied from (approximately) 1 in 10^3 RNA copies being infectious to as low as 1 in 10^7 . Based on samples collected over several days post-infection, Lin et al. [59] found a lognormal distribution for the RNA to pfu ratio with a mean of 1.6×10^5 . They additionally concluded that the day of symptom onset would represent the highest infectivity values. Recently Kitagawa et al. [54] assessed the concentration of virus RNA and viable virus concentration in hospital room air occupied by SARS-CoV-2 patients and found a direct correlation between these quantities. Their reported results for infectivity fell within the range of values observed in the studies mentioned before. These observations emphasize the high variance in infectivity values and how it contradicts the use of a singular dose-response constant value. Sender et al. [93] detailed possible reasons for these uncertainties across available literature including assays used for measuring infectious titer not having optimal conditions for SARS-CoV-2 infection and the possibility of immune response effect showing up in the ratio. These discrepancies were also highlighted by McCormick and Mermel [69] where they compared the virus particle to pfu ratio across different viruses and pointed out that certain viruses with lower infectivity on record have higher reproduction number and associated this contradiction to possible suboptimal conditions for tissue culture of certain viruses along with a lack of standardized means of generating virus particle to pfu values, which would otherwise allow for comparison between the various values reported across publications. In general, the quantity of viable virus in air would have dependency on a variety of factors—type of respiratory event, use of masks, environmental conditions, virus variant, and viral load, thus suggesting that to obtain a range of infectivity with a strong confidence interval, a lot more work remains to be done. This also ties back to the aerosol transmission models where if spatial variation in virus particles is not taken into account, every susceptible might be falling on one side of the infectivity threshold and thus we would be underestimating (or overestimating) the number of secondary infections.

The second factor (other than infectivity) that determines the dose-response constant for viral RNA in air is the dose-response constant based on viable infectious

virus. The latest data for controlled inoculation (of SARS-CoV-2) in humans was provided by Killingley et al. [53]. They inoculated 36 volunteers with (a dose of) 10TCID₅₀ of the wild variant SARS-CoV-2 and kept them under close observation, based on which they reported infection appearing in 53% of individuals. Replacing the data from this model in Haas' [42] dose-response model would provide a dose-response parameter value. As new data arises for a virus, this quantity can be updated and improved in tandem.

After addressing all uncertainties to the best of one's knowledge, one would be left with a dose-response model of choice that can be practically applied to relate the infectious aerosols to the final endpoint effect of illness. A complete model that encapsulates the entire process of transmission is provided in the next section in an attempt to obtain a distribution for the number of secondary infections.

6.7 Probability Distribution for Number of Secondary Infections Z

The aim of this model (detailed in Chaudhuri et al. [17]) is to generate a distribution for the number of secondary infections due to singular index cases inside a collection of indoor locations while accounting for the underlying physics yet retaining simplicity to make it practically viable. To obtain a distribution, we will be solving the model to obtain one sample of secondary infection count in each iteration of a Monte Carlo simulation. Each iteration will need to capture all four parts of the airborne transmission process as discussed in this chapter.

The first stage is the viral load distribution which is provided by Yang et al. [110] that captures the quantity of infectious matter within the index case. This has been shown to attain a lognormal form [110]

$$f(\rho_v) = \frac{1}{\rho_v \sigma \sqrt{2\pi}} e^{-(\ln(\rho_v) - \mu)^2 / 2\sigma^2} \quad (6.3)$$

where ρ_v is the random variable corresponding to the viral load ρ . To describe the size distribution of the aerosols within which this viral matter resides, we apply the n -mode PSD described in Eq. 6.1 [81]. These descriptions are combined by integrating the volume size distribution (obtained from the PSD) over the entire range of particle diameters and then multiplying with the viral load and volume flow rate of ejected airflow [81], to obtain the continuous source term, i.e., the total number of virions ejected per unit time Q_{x_0, y_0, z_0} by the index case located at the spatial coordinates (x_0, y_0, z_0) . This acts as an input to the model that will describe the aerosol transmission and connect the index case to the susceptible. Because we are looking at indoor locations over a timescale that far exceeds the transitory puff regime, a justification can be made to apply a purely diffusive method—which in our case is the Drivas-Cheng model [19, 29], given by

$$\frac{\partial c}{\partial t} = D_T \left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} + \frac{\partial^2 c}{\partial z^2} \right) - \left(a + \frac{w_d A}{V} \right) c \quad (6.4)$$

where c is the concentration of virions in air, a is the ventilation rate, A is the room area, V is the room volume, w_d is the deposition velocity, and D_T is the turbulent diffusivity which was expressed as an empirical function of a and A by Cheng et al. [19]. The solution to this equation was provided by Drivas et al. [29] in the form

$$c(x, y, z, t') = \int_0^{t'} \frac{Q_{x_0, y_0, z_0} e^{-\left(a + \frac{w_d A}{V}\right)t}}{(4\pi D_T t)^{3/2}} R_x R_y R_z \psi(t) dt \quad (6.5)$$

where R corresponds to the wall reflection terms that attain algebraic forms [29]. The term $\psi(t)$ quantifies the virus survivability function. The concentration field c now provides us with the virion concentration that any susceptible is exposed to at the spatial coordinates (x, y, z) . This can be converted to the number of virions inhaled, by relating it with the susceptible's inhalation rate \dot{V}_b over an exposure duration τ , as

$$N_v(x, y, z, \tau) = \int_0^\tau c(x, y, z, t') \dot{V}_b dt' \quad (6.6)$$

The reaction of the susceptible's biology to the incoming infectious matter N_v is modeled using Haas' [41, 42] dose-response model with the dose-response constant r_v as input. This gives us the probability of infection \mathcal{P} for an individual as described by Eq. 6.2, which can be converted to the number of secondary infections Z at a location with population density ρ_p , over an exposure duration of τ , through an integration operation as follows:

$$Z = \int \int \rho_p \mathcal{P}(x, y, z, \tau) dx dy \quad (6.7)$$

The population density term is assumed to be uniform, thus relegating Eq. 6.7 to an area averaging operation on the probability of infection field. The quantity Z is generated in each iteration of the simulation (each corresponding to one indoor location), and the final set of values is used to obtain the distribution shown in Fig. 6.7. Even though this approach provides a model that encapsulates the entire process, an analytical formulation for the pdf of Z would be more practically applicable. To this end, we would require further simplifications in our approach to make the problem analytically tractable. Thus, in this method we will be assuming well-mixed conditions within the indoor location. Results from the simulations (shown in Chaudhuri et al. [17]) highlight the viral load ρ as one of the most dominant parameters in determining Z . Based on this observation, the logical follow-up is to determine a direct relation between the viral load and the number of secondary infections, which is done by first performing a normalization and defining

the secondary attack rate as $\tilde{Z} = Z/n$, where n corresponds to the occupancy of an indoor location. The observed variation of \tilde{Z} with respect to ρ , as seen in the simulations performed by Chaudhuri et al. [17], mimics the behavior of the dose-response model suggesting that the primary contribution to N_v in the dose-response equation (Eq. 6.2) is indeed from the viral load. Hence, the following relation is proposed (similar to the dose-response relation):

$$\tilde{Z} = 1 - e^{-\alpha\rho} \quad (6.8)$$

where α is a constant, which can be derived by employing the well-mixed conditions along with appropriate known input parameters, to give

$$\alpha = \frac{r_v \langle \dot{Q}_l \rangle \langle t_s \rangle \dot{V}_b}{\langle V \rangle} \int_0^{\langle \tau \rangle} \psi(t) e^{-(\langle a \rangle / 3600 + \beta_0)t} dt \quad (6.9)$$

Here, \dot{Q}_l is the ejected liquid volume per unit time, t_s is the duration of the ejection event, and $\beta_0 = w_d A / V$ is the wall deposition coefficient. The operator $\langle \rangle$ denotes the ensemble averaging operation. At this stage, one can combine Eq. 6.8, the viral load distribution in Eq. 6.3, and the relation $\tilde{Z} = Z/n$ to write the analytical pdf $\phi(\tilde{Z})$ (Z , \tilde{Z} and N are random variables corresponding to Z , \tilde{Z} and n , respectively). This is expressed as

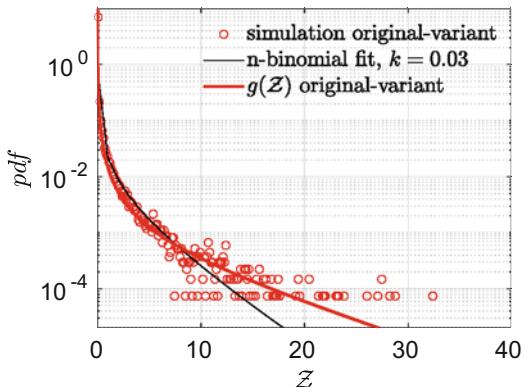
$$\phi(\tilde{Z}) = f(\rho_v) \frac{d\rho_v}{d\tilde{Z}} = f\left(-\frac{1}{\alpha} \ln(1 - \tilde{Z})\right) \frac{d\left(-\frac{1}{\alpha} \ln(1 - \tilde{Z})\right)}{d\tilde{Z}} \quad (6.10)$$

Now, given that the distribution of N across all the indoor locations (from occupancy data) is known as $h(N)$, using the formulation for the pdf of the product of independent random variables (N and \tilde{Z}), we finally obtain the pdf for the number of secondary infections

$$g(Z) = \int_0^\infty \frac{h(N)}{-\sigma \sqrt{2\pi} (N - Z) \ln(1 - Z/N)} e^{-\left\{ \ln\left(-\frac{1}{\alpha} \ln(1 - Z/N)\right) - \mu \right\}^2 / 2\sigma^2} dN \quad (6.11)$$

where μ and σ are the parameters describing the lognormal viral load distribution. The comparison between the analytical description of $g(Z)$ and its simulated variant is shown in Fig. 6.7 which shows clear overlap between the curves, hence presenting a strong case for simplified models when looking at airborne disease spread dynamics through a statistical lens. The curves are long-tailed and describe the overdispersed nature of SARS-CoV-2 emphasizing the existence of superspreading events where a large number of infections occur due to a single index case. For further details on the process along with effect of mitigation measures and how they impact the pdf of Z , the reader can read the study by Chaudhuri et al. [17]. The primary uncertainty within this description lies in the constant value of r_v that has

Fig. 6.7 Pdf of the number of secondary infections: \mathcal{Z} calculated from Monte Carlo simulations and an analytical formulation (described by Eq. 6.11), along with a fitted negative binomial pdf. Reprinted from Chaudhuri et al. [17], with the permission of AIP Publishing



been taken from Schijven et al. [92] as has been addressed in Sect. 6.6. Immediate future work on this model involves validation with a real-life scenario followed by additional introduction of complexities based on the validation exercise and then finally using the pdf of secondary infections to obtain an estimate for the pdf of reproduction number.

6.8 Conclusion

In this chapter we have looked at the airborne virus transmission process in the context of a statistical description of an outbreak and broken it down into sub-topics where each has its own complexities and uncertainties in understanding and modeling. We started by looking at the density of the virus particles in the body of the infectious individual and how the viral load is quantified and measured from real-life cases. Virus particles are embedded within the ejected aerosols whose physics strongly depend on their size distribution. Studies on such distributions have shown variations across the board, some owing to measurement techniques while others due to biological and environmental factors. The current models that best represent such distributions have been described. The succeeding discussion on the various treatments of aerosol transmission from the infectious individual to the susceptible revealed the different levels of complexity involved in modeling this process and poses the question of how one should choose the appropriate model depending on the end goal of a study. A need for intensive simulations of the transmission process exists for proper understanding of the underlying physics which can then guide modeling principles while also serving as a reference for comparison of simpler models that are more practical to use. The discussion on the human body's response to the incoming virus-laden aerosols shed light on ways to connect the aerosol physics to the biological aspects of airborne transmission. Some of the uncertainties associated with the simpler well-mixed models are enhanced because of the localized virion concentration being an important factor to account

for due to the high variance in reported virus infectivity values. Finally, we conclude with a discussion of a model that can capture the entire transmission process to eventually provide a distribution for the number of secondary infections and paves the way for developing method that can accurately predict virus spread.

References

1. Abkarian, M., Stone, H.A.: Stretching and break-up of saliva filaments during speech: A route for pathogen aerosolization and its potential mitigation. *Phys. Rev. Fluids* **5**, 102301 (2020). <https://doi.org/10.1103/PhysRevFluids.5.102301>
2. Almstrand, A.C., Bake, B., Ljungström, E., Larsson, P., Bredberg, A., Mirgorodskaya, E., Olin, A.C.: Effect of airway opening on production of exhaled particles. *J. Appl. Physiol.* **108**(3), 584–588 (2010). <https://doi.org/10.1152/japplphysiol.00873.2009>. PMID: 20056850
3. Alsved, M., Holm, S., Christiansen, S., Smidt, M., Rosati, B., Ling, M., Boesen, T., Finster, K., Bilde, M., Löndahl, J., et al.: Effect of aerosolization and drying on the viability of pseudomonas syringae cells. *Front. Microbiol.* **9**, 3086 (2018)
4. Alsved, M., Matamis, A., Bohlin, R., Richter, M., Bengtsson, P.E., Fraenkel, C.J., Medstrand, P., Löndahl, J.: Exhaled respiratory particles during singing and talking. *Aerosol Sci. Technol.* **54**(11), 1245–1248 (2020)
5. Anand, S., Mayya, Y.: Size distribution of virus laden droplets from expiratory ejecta of infected subjects. *Sci. Rep.* **10**, 21174 (2020)
6. Asadi, S., Wexler, A., Cappa, C., et al.: Aerosol emission and superemission during human speech increase with voice loudness. *Sci. Rep.* **9**, 2348 (2019)
7. Balachandar, S., Zaleski, S., Soldati, A., Ahmadi, G., Bourouiba, L.: Host-to-host airborne transmission as a multiphase flow problem for science-based social distance guidelines. *Int. J. Multiphase Flow* **132**, 103439 (2020)
8. Bhavnani, D., James, E., Johnson, K., et al.: SARS-CoV-2 viral load is associated with risk of transmission to household and community contacts. *BMC Infect. Dis.* **22**, 672 (2022)
9. Bi, P., Wang, J., Hiller, J.E.: Weather: driving force behind the transmission of severe acute respiratory syndrome in China? *Internal Med. J.* **37**(8), 550–554
10. Bourouiba, L.: The fluid dynamics of disease transmission. *Ann. Rev. Fluid Mech.* **53**, 473–508 (2021)
11. Bourouiba, L., Dehandschoewercker, E., Bush, J.W.: Violent expiratory events: on coughing and sneezing. *J. Fluid Mech.* **745**, 537–563 (2014)
12. Álvaro Briz-Redón, Ángel Serrano-Aroca: a spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. *Sci. Total Environ.* **728**, 138811 (2020)
13. Challenger, J., Foo, C., Wu, Y., et al.: Modelling upper respiratory viral load dynamics of SARS-CoV-2. *BMC Med.* **20**, 25 (2022)
14. Chao, C.Y.H., Wan, M.P., Morawska, L., Johnson, G.R., Ristovski, Z., Hargreaves, M., Mengersen, K., Corbett, S., Li, Y., Xie, X., et al.: Characterization of expiration air jets and droplet size distributions immediately at the mouth opening. *J. Aerosol Sci.* **40**(2), 122–133 (2009)
15. Chaudhuri, S., Basu, S., Kabi, P., Unni, V.R., Saha, A.: Modeling the role of respiratory droplets in covid-19 type pandemics. *Phys. Fluids* **32**(6), 063309 (2020)
16. Chaudhuri, S., Basu, S., Saha, A.: Analyzing the dominant SARS-CoV-2 transmission routes toward an ab initio disease spread model. *Phys. Fluids* **32**(12), 123306 (2020)
17. Chaudhuri, S., Kasibhatla, P., Mukherjee, A., Pan, W., Morrison, G., Mishra, S., Murty, V.K.: Analysis of overdispersion in airborne transmission of COVID-19. *Phys. Fluids* **34**, 051914 (2022)

18. Chen, K.M., Jiang, X., Kimerling, L.C., Hammond, P.T.: Selective self-organization of colloids on patterned polyelectrolyte templates. *Langmuir* **16**(20), 7825–7834 (2000)
19. Cheng, K.C., Acevedo-Bolton, V., Jiang, R.T., Klepeis, N.E., Ott, W.R., Fringer, O.B., Hildemann, L.M.: Modeling exposure close to air pollution sources in naturally ventilated residences: association of turbulent diffusion coefficient with air change rate. *Environ. Sci. Technol.* **45**(9), 4016–4022 (2011)
20. Chong, K.L., Ng, C.S., Hori, N., Yang, R., Verzicco, R., Lohse, D.: Extended lifetime of respiratory droplets in a turbulent vapor puff and its implications on airborne disease transmission. *Phys. Rev. Lett.* **126**(3), 034502 (2021)
21. Coleman, K.K., Tay, D.J.W., Tan, K.S., et al.: Viral load of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in respiratory aerosols emitted by patients with coronavirus disease 2019 (COVID-19) while breathing, talking, and singing. *Clin. Infect. Dis.* **74**(10), 1722–1728 (2021)
22. Cowling, B., Ip, D., Fang, V., et al.: Aerosol transmission is an important mode of influenza A virus spread. *Nat. Commun.* **4**(1935) (2013)
23. Dbouk, T., Drikakis, D.: On coughing and airborne droplet transmission to humans. *Phys. Fluids* **32**(5), 053310 (2020)
24. Deegan, R.D., Bakajin, O., Dupont, T.F., Huber, G., Nagel, S.R., Witten, T.A.: Capillary flow as the cause of ring stains from dried liquid drops. *Nature* **389**(6653), 827–829 (1997)
25. Despres, H.W., Mills, M.G., Shirley, D.J., Schmidt, M.M., Huang, M.L., Roychoudhury, P., Jerome, K.R., Greninger, A.L., Bruce, E.A.: Measuring infectious SARS-CoV-2 in clinical samples reveals a higher viral titer:rna ratio for delta and epsilon vs. alpha variants. *Proc. Natl. Acad. Sci.* **119**(5), e2116518119 (2022)
26. Devineau, S., Anyfantakis, M., Marichal, L., Kiger, L., Morel, M., Rudiuk, S., Baigl, D.: Protein adsorption and reorganization on nanoparticles probed by the coffee-ring effect: application to single point mutation detection. *J. Am. Chem. Soc.* **138**(36), 11623–11632 (2016)
27. Dhand, R., Li, J.: Coughs and sneezes: their role in transmission of respiratory viral infections, including SARS-CoV-2. *Am. J. Respir. Crit. Care Med.* **202**(5), 651–659 (2020)
28. Di Martino, P., Cafferini, N., Joly, B., Darfeuille-Michaud, A.: Klebsiella pneumoniae type 3 pili facilitate adherence and biofilm formation on abiotic surfaces. *Res. Microbiol.* **154**(1), 9–16 (2003)
29. Drivas, P.J., Valberg, P.A., Murphy, B.L., Wilson, R.: Modeling indoor air exposure from short-term point source releases. *Indoor Air* **6**(4), 271–277 (1996)
30. Duguid, J.: The numbers and the sites of origin of the droplets expelled during expiratory activities. *Edinb. Med. J.* **52**(11), 385 (1945)
31. Duguid, J.: The size and the duration of air-carriage of respiratory droplets and droplet-nuclei. *Epidemiol. Infect.* **44**(6), 471–479 (1946)
32. Efstratiou, M., Christy, J., Sefiane, K.: Crystallization-driven flows within evaporating aqueous saline droplets. *Langmuir* **36**(18), 4995–5002 (2020)
33. Eslami, H., Jalili, M.: The role of environmental factors to transmission of SARS-CoV-2 (COVID-19). *AMB Express* **10**(1) (2020)
34. Euser, S., Aronson, S., Manders, I., et al.: SARS-CoV-2 viral-load distribution reveals that viral loads increase with age: a retrospective cross-sectional cohort study. *Int. J. Epidemiol.* **50**(6) (2022)
35. Fabregat, A., Gisbert, F., Vernet, A., Dutta, S., Mittal, K., Pallarès, J.: Direct numerical simulation of the turbulent flow generated during a violent expiratory event. *Phys. Fluids* **33**(3), 035122 (2021)
36. Feng, Y., Marchal, T., Sperry, T., Yi, H.: Influence of wind and relative humidity on the social distancing effectiveness to prevent COVID-19 airborne transmission: a numerical study. *J. Aerosol. Sci.* **147**, 105585 (2020)
37. Giri, A., Dutta Choudhury, M., Dutta, T., Tarafdar, S.: Multifractal growth of crystalline NaCl aggregates in a gelatin medium. *Crystal Growth Design* **13**(1), 341–345 (2013)

38. Gorr, H.M., Zueger, J.M., McAdams, D.R., Barnard, J.A.: Salt-induced pattern formation in evaporating droplets of lysozyme solutions. *Colloids Surf. B: Biointerfaces* **103**, 59–66 (2013)
39. Greenhalgh, T., Jimenez, J.L., Prather, K.A., Tufekci, Z., Fisman, D., Schooley, R.: Ten scientific reasons in support of airborne transmission of SARS-CoV-2. *Lancet* **397**(10285), 1603–1605 (2021)
40. Gregson, F.K.A., Watson, N.A., Orton, C.M., Haddrell, A.E., McCarthy, L.P., Finnie, T.J.R., Gent, N., Donaldson, G.C., Shah, P.L., Calder, J.D., Bzdek, B.R., Costello, D., Reid, J.P.: Comparing aerosol concentrations and particle size distributions generated by singing, speaking and breathing. *Aerosol Sci. Technol.* **55**(6), 681–691 (2021)
41. HAAS, C.N.: Estimation of risk due to low doses of microorganisms: a comparison of alternative methodologies. *Am. J. Epidemiol.* **118**(4), 573–582 (1983)
42. Haas, C.N.: Action levels for SARS-CoV-2 in air: preliminary approach. *Risk Anal.* **41**(5), 705–709 (2021)
43. Han, Z., Weng, W., Huang, Q.: Characterizations of particle size distribution of the droplets exhaled by sneeze. *J. R. Soc. Interface* **10**(88), 20130560 (2013)
44. Hathway, E., Noakes, C., Sleigh, P., Fletcher, L.: CFD simulation of airborne pathogen transport due to human activities. *Build. Environ.* **46**(12), 2500–2511 (2011)
45. Department of Homeland Security, U.S.: Estimated airborne decay of SARS-CoV-2. <https://www.dhs.gov/science-and-technology/sars-airborne-calculator>. Tech. rep.
46. Hu, M., Lin, H., Wang, J., Xu, C., Tatem, A.J., Meng, B., Zhang, X., Liu, Y., Wang, P., Wu, G., Xie, H., Lai, S.: Risk of Coronavirus Disease 2019 Transmission in train passengers: an epidemiological and modeling study. *Clin. Infect. Dis.* **72**(4), 604–610 (2020)
47. Johnson, G., Morawska, L., Ristovski, Z., Hargreaves, M., Mengersen, K., Chao, C., Wan, M., Li, Y., Xie, X., Katoshevski, D., Corbett, S.: Modality of human expired aerosol size distributions. *J. Aerosol Sci.* **42**(12), 839–851 (2011). <https://doi.org/10.1016/j.jaerosci.2011.07.009>
48. Johnson, G.R., Morawska, L.: The mechanism of breath aerosol formation. *J. Aerosol Med. Pulm. Drug Delivery* **22**(3), 229–237 (2009)
49. Jüni, P., Baert, S., Bobos et al., P.: Rapid antigen tests for voluntary screen testing. *Science Briefs Ontario COVID-19 Science Advisory Table* **2**, 52 (2021)
50. Kang, M., Xin, H., Yuan, L., et al.: Transmission dynamics and epidemiological characteristics of delta variant infections in China. *medRxiv* (2021)
51. Kawasaji, H., Takegoshi, Y., Kaneda, M., Ueno, A., Miyajima, Y., Kawago, K., Fukui, Y., Yoshida, Y., Kimura, M., Yamada, H., et al.: Transmissibility of COVID-19 depends on the viral load around onset in adult and symptomatic patients. *PLoS One* **15**(12), e0243597 (2020)
52. Keil, C.B.: A tiered approach to deterministic models for indoor air exposures. *Appl. Occup. Environ. Hyg.* **15**(1), 145–151 (2000). <https://doi.org/10.1080/104732200301962>. PMID: 10712069
53. Killingley, B., Mann, A.J., Kalinova, M., et al.: Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge in young adults. *Nat. Med.* **28**, 1031–1041 (2022). <https://doi.org/10.1038/s41591-022-01780-9>
54. Kitagawa, H., Nomura, T., Kaiki, Y., Kakimoto, M., Nazmul, T., Omori, K., Shigemoto, N., Sakaguchi, T., Ohge, H.: Viable SARS-CoV-2 detected in the air of hospital rooms of COVID-19 patients with early infection. *Int. J. Infect. Dis.* (2022)
55. Kramer, A., Schwebke, I., Kampf, G.: How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. *BMC Infect. Dis.* **6**(1), 1–8 (2006)
56. Li, B., Deng, A., Li, K., Hu, Y., et al.: Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 delta variant. *medRxiv* (2021). <https://doi.org/10.1101/2021.07.07.21260122>
57. Li, H., Leong, F.Y., Xu, G., Ge, Z., Kang, C.W., Lim, K.H.: Dispersion of evaporating cough droplets in tropical outdoor environment. *Phys. Fluids* **32**(11), 113301 (2020)
58. Lin, G., Hamilton, A., Gatalo, O., et al.: Investigating the effects of absolute humidity and movement on COVID-19 seasonality in the united states. *Sci. Rep.* **12**(16729) (2022)

59. Lin, Y., Malott, R., Ward, L., et al.: Detection and quantification of infectious severe acute respiratory coronavirus-2 in diverse clinical and environmental samples. *Sci. Rep.* **12**(5418) (2022)
60. Loudon, R., Roberts, R.: Droplet expulsion from the respiratory tract. *Am. Rev. Respir. Dis.* **95**(3), 435–42 (1967)
61. Lucey, M., Macori, G., Mullane, N., Sutton-Fitzpatrick, U., Gonzalez, G., Coughlan, S., Purcell, A., Fenelon, L., Fanning, S., Schaffer, K.: Whole-genome sequencing to track severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission in nosocomial outbreaks. *Clin. Infect. Dis.* **72**(11), e727–e735 (2020)
62. Lustig, S.R., Biswakarma, J.J., Rana, D., Tilford, S.H., Hu, W., Su, M., Rosenblatt, M.S.: Effectiveness of common fabrics to block aqueous aerosols of virus-like nanoparticles. *ACS Nano* **14**(6), 7651–7658 (2020)
63. Maheshwari, S., Zhang, L., Zhu, Y., Chang, H.C.: Coupling between precipitation and contact-line dynamics: multiring stains and stick-slip motion. *Phys. Rev. Lett.* **100**(4), 044503 (2008)
64. Majee, S., Chowdhury, A.R., Pinto, R., Chattopadhyay, A., Agharkar, A.N., Chakravorty, D., Basu, S.: Spatiotemporal evaporating droplet dynamics on fomites enhances long term bacterial pathogenesis. *Commun. Biol.* **4**(1), 1–16 (2021)
65. Marin, A., Karpitschka, S., Noguera-Marín, D., Cabrerizo-Vilchez, M.A., Rossi, M., Kähler, C.J., Valverde, M.A.R.: Solutal marangoni flow as the cause of ring stains from drying salty colloidal drops. *Phys. Rev. Fluids* **4**(4), 041601 (2019)
66. Marin, A.G., Gelderblom, H., Lohse, D., Snoeijer, J.H.: Order-to-disorder transition in ring-shaped colloidal stains. *Phys. Rev. Lett.* **107**(8), 085502 (2011)
67. Marr, L.C., Tang, J.W., Van Mullekom, J., Lakdawala, S.S.: Mechanistic insights into the effect of humidity on airborne influenza virus survival, transmission and incidence. *J. R. Soc. Interface* **16**(150), 20180298 (2019)
68. Martin, D., Noakes, R.: Crystal settling in a vigorously convecting magma chamber. *Nature* **332**, 534–536 (1988)
69. McCormick, W., Mermel, L.: The basic reproductive number and particle-to-plaque ratio: comparison of these two parameters of viral infectivity. *Virol. J.* **18**(92) (2021)
70. Mondelli, M.U., Colaneri, M., Seminari, E.M., Baldanti, F., Bruno, R.: Low risk of SARS-CoV-2 transmission by fomites in real-life conditions. *Lancet Infect. Dis.* **21**(5), e112 (2021). [https://doi.org/10.1016/S1473-3099\(20\)30678-2](https://doi.org/10.1016/S1473-3099(20)30678-2)
71. Moran, P.: The dilution assay of viruses. *J. Hyg.* **52**(2), 189–193 (1954)
72. Morawska, L., Milton, D.K.: It is time to address airborne transmission of COVID-19. *Clin. Infect. Dis.* (2020). <https://doi.org/10.1093/cid/ciaa939>
73. Morton, B.R., Taylor, G., Turner, J.S.: Turbulent gravitational convection from maintained and instantaneous sources. *Proc. R. Soc. Lond. A* **234**(1196), 1–23 (1956). <https://doi.org/10.1098/rspa.1956.0011>
74. Nicas, M., Nazaroff, W.W., Hubbard, A.: Toward understanding the risk of secondary airborne infection: emission of respirable pathogens. *J. Occup. Environ. Hyg.* **2**(3), 143–154 (2005)
75. Organization, W.H., et al.: Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations: scientific brief, 27 March 2020. Tech. rep., World Health Organization (2020)
76. Organization, W.H., et al.: Transmission of SARS-CoV-2: implications for infection prevention precautions, scientific brief. 9 July 2020. Tech. rep., World Health Organization (2020)
77. Özkaynak, H., Ryan, P., Allen, G., Turner, W.: Indoor air quality modeling: compartmental approach with reactive chemistry. *Environ. Int.* **8**(1), 461–471 (1982). *Indoor Air Pollution*
78. Papineni, R.S., Rosenthal, F.S.: The size distribution of droplets in the exhaled breath of healthy human subjects. *J. Aerosol Med.* **10**(2), 105–116 (1997)
79. Paredi, P., Kharitonov, S.A., Barnes, P.J.: Analysis of expired air for oxidation products. *Am. J. Respir. Crit. Care Med.* **166**, 31S–37S (2002)
80. Pathak, B., Christy, J., Sefiane, K., Gozuacik, D.: Complex pattern formation in solutions of protein and mixed salts using dehydrating sessile droplets. *Langmuir* **36**(33), 9728–9737 (2020)

81. Pöhlker, M.L., Krüger, O.O., Förster, J.D., Berkemeier, T., Elbert, W., Fröhlich-Nowoisky, J., Pöschl, U., Pöhlker, C., Bagheri, G., Bodenschatz, E., et al.: Respiratory aerosols and droplets in the transmission of infectious diseases (2021). arXiv preprint arXiv:2103.01188
82. Prather, K.A., Marr, L.C., Schooley, R.T., McDiarmid, M.A., Wilson, M.E., Milton, D.K.: Airborne transmission of SARS-CoV-2. *Science* **370**(6514), 303–304 (2020)
83. Public Health England: Understanding cycle threshold (ct) in SARS-CoV-2 RT-PCR (2020). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/926410/Understanding_Cycle_Threshold__Ct_in_SARS-CoV-2_RT-PCR_.pdf
84. Public Health Ontario: An overview of cycle threshold values and their role in SARS-CoV-2 real-time PCR test interpretation (2020). <https://www.publichealthontario.ca/-/media/documents/nkov/main/2020/09/cycle-threshold-values-sars-cov2-pcr.pdf?la=en>
85. Rasheed, A., Hegde, O., Chatterjee, R., Sampathirao, S.R., Chakravortty, D., Basu, S.: Physics of self-assembly and morpho-topological changes of klebsiella pneumoniae in desiccating sessile droplets. *J. Colloid Interface Sci.* **629**, 620–631 (2023)
86. Rasheed, A., Sharma, S., Kabi, P., Saha, A., Chaudhuri, S., Basu, S.: Precipitation dynamics of surrogate respiratory sessile droplets leading to possible fomites. *J. Colloid Interface Sci.* **600**, 1–13 (2021)
87. Richardson, L.F., Walker, G.T.: Atmospheric diffusion shown on a distance-neighbour graph. *Proc. R. Soc. Lond. A: Containing Papers Math. Phys. Charact.* **110**(756), 709–737 (1926). <https://doi.org/10.1098/rspa.1926.0043>
88. Riley, E., Murphy, G., Riley, R.: Airborne spread of measles in a suburban elementary school. *Am. J. Epidemiol.* **107**(5), 421–432 (1978)
89. Rosti, M.E., Cavaiola, M., Olivieri, S., Seminara, A., Mazzino, A.: Turbulence role in the fate of virus-containing droplets in violent expiratory events. *Phys. Rev. Res.* **3**, 013091 (2021)
90. Rowan, S., Newton, M., Driewer, F., McHale, G.: Evaporation of microdroplets of azeotropic liquids. *J. Phys. Chem. B* **104**(34), 8217–8220 (2000)
91. Ryan, P., Spengler, J., Halfpenny, P.: Sequential box models for indoor air quality: application to airliner cabin air quality. *Atmos. Environ.* (1967) **22**(6), 1031–1038 (1988)
92. Schijven, J., Vermeulen, L.C., Swart, A., Meijer, A., Duizer, E., de Roda Husman, A.M.: Quantitative microbial risk assessment for airborne transmission of SARS-CoV-2 via breathing, speaking, singing, coughing, and sneezing. *Environ. Health Perspect.* **129**(4), 047002 (2021)
93. Sender, R., Bar-On, Y.M., Gleizer, S., Bernshtain, B., Flamholz, A., Phillips, R., Milo, R.: The total number and mass of SARS-CoV-2 virions. *Proc. Natl. Acad. Sci.* **118**(25), e2024815118 (2021)
94. Sett, A., Ayushman, M., Dasgupta, S., DasGupta, S.: Analysis of the distinct pattern formation of globular proteins in the presence of micro-and nanoparticles. *J. Phys. Chem. B* **122**(38), 8972–8984 (2018)
95. Sharma, S., Pinto, R., Saha, A., Chaudhuri, S., Basu, S.: On secondary atomization and blockage of surrogate cough droplets in single-and multilayer face masks. *Sci. Adv.* **7**(10), eabf0452 (2021)
96. Smithier, S.J., Lear-Rooney, C., Biggins, J., Pettitt, J., Lever, M.S., Olinger, G.G.: Comparison of the plaque assay and 50% tissue culture infectious dose assay as methods for measuring filovirus infectivity. *J. Virol. Methods* **193**(2), 565–571 (2013). <https://doi.org/10.1016/j.jviromet.2013.05.015>
97. Sobral, M.F.F., Duarte, G.B., da Penha Sobral, A.I.G., Marinho, M.L.M., de Souza Melo, A.: Association between climate variables and global transmission of SARS-CoV-2. *Sci. Total Environ.* **729**, 138997 (2020)
98. Sze To, G., Chao, C.: Review and comparison between the Wells–Riley and dose-response approaches to risk assessment of infectious respiratory diseases. *Indoor Air* **20**(1), 2–16 (2010)
99. Tan, J., Mu, L., Huang, J., Yu, S., Chen, B., Yin, J.: An initial investigation of the association between the SARS outbreak and weather: with the view of the environmental temperature and its variation. *J. Epidemiol. Community Health* **59**(3), 186–192 (2005)

100. To, T., Zhang, K., Maguire, B., Terebessy, E., Fong, I., Parikh, S., Zhu, J.: Correlation of ambient temperature and COVID-19 incidence in Canada. *Sci. Total Environ.* **750**, 141484 (2021)
101. Van Loosdrecht, M., Lyklema, J., Norde, W., Schraa, G., Zehnder, A.: The role of bacterial cell wall hydrophobicity in adhesion. *Appl. Environ. Microbiol.* **53**(8), 1893–1897 (1987)
102. Vejerano, E.P., Marr, L.C.: Physico-chemical characteristics of evaporating respiratory fluid droplets. *J. R. Soc. Interface* **15**(139), 20170939 (2018)
103. Venkatram, A., Weil, J.: Modeling turbulent transport of aerosols inside rooms using eddy diffusivity. *Indoor air* **31**(18) (2021)
104. Watanabe, T., Bartrand, T., Weir, M., Omura, T., Haas, C.: Development of a dose-response model for SARS coronavirus. *Risk Anal.* **30**(7), 1129–1138 (2010)
105. Wells, W.: On air-borne infection: Study ii. droplets and droplet nuclei. *Am. J. Epidemiol.* **20**(3), 611–618 (1934)
106. Wells, W.F., et al.: Airborne contagion and air hygiene. an ecological study of droplet infections. *Airborne Contagion and Air Hygiene. An Ecological Study of Droplet Infections* (1955)
107. Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., Liu, M.: Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *Sci. Total Environ.* **729**, 139051 (2020). <https://doi.org/10.1016/j.scitotenv.2020.139051>
108. Xie, X., Li, Y., Sun, H., Liu, L.: Exhaled droplets due to talking and coughing. *J. R. Soc. Interface* **6**(suppl_6), S703–S714 (2009)
109. Xie, X., Li, Y., Zhang, T., Fang, H.H.: Bacterial survival in evaporating deposited droplets on a teflon-coated surface. *Appl. Microbiol. Biotechnol.* **73**(3), 703–712 (2006)
110. Yang, Q., Saldi, T.K., Gonzales, P.K., Lasda, E., Decker, C.J., Tat, K.L., Fink, M.R., Hager, C.R., Davis, J.C., Ozeroff, C.D., et al.: Just 2% of SARS-CoV-2- positive individuals carry 90% of the virus circulating in communities. *Proc. Natl. Acad. Sci.* **118**(21) (2021). <https://doi.org/10.1073/pnas.2104547118>
111. Yunker, P.J., Still, T., Lohr, M.A., Yodh, A.: Suppression of the coffee-ring effect by shape-dependent capillary interactions. *Nature* **476**(7360), 308–311 (2011)

Chapter 7

Modeling Mutation-Driven Emergence of Drug-Resistance: A Case Study of SARS-CoV-2



Congjie Shi, Thomas N. Vilches, Ao Li, Jianhong Wu,
and Seyed M. Moghadas

7.1 Introduction

Emergence of drug-resistance (DR) remains a major challenge in the control or elimination of several infectious diseases [2, 4, 28, 30, 33]. In particular, pathogens with high mutation rates could naturally acquire resistant mutants, which may become widespread in the presence of treatment and strong selection pressure [7]. There is a vast literature on the mechanisms of antimicrobial resistance for bacterial diseases and antiviral resistance for viral infections. Various strategies have been deployed to reduce the rate of DR emergence and spread, including drug stewardship and adherence to treatment regimens, combination therapy, the development of novel potent drugs, and expansion of preventive measures (e.g., vaccination) [8, 17, 33, 43].

The effect of DR emergence and population-wide spread has been evaluated in numerous modeling studies [1, 3, 18, 25, 29, 35]. The structure of most models encapsulates the dynamics of DR explicitly arising from the pressure exerted during treatment. Although this selection pressure may encourage the evolution of resistant mutants and their subsequent spread, the presence of such mutants before the start of treatment remains critically important to the emergence of DR [29]. Here, we develop a mutation-driven model to evaluate the effect of naturally evolving mutants prior to the start of treatment on the development of de novo DR after the start of treatment. We construct our model to represent the dynamics of disease

C. Shi · T. N. Vilches · A. Li · S. M. Moghadas (✉)
Agent-Based Modelling Laboratory, York University, Toronto, ON, Canada
e-mail: moghadas@yorku.ca

J. Wu
Laboratory for Industrial and Applied Mathematics, Department of Mathematics and Statistics,
York University, Toronto, ON, Canada
e-mail: wujh@yorku.ca

transmission based on the natural history of SARS-CoV-2 that has caused the COVID-19 pandemic with significant health and socioeconomic burden globally.

It is well-documented that SARS-CoV-2 exhibits high mutation rates both in-host and inter-host [39]. Some mutations are associated with higher transmissibility, and a larger escape from neutralizing antibodies acquired naturally or through vaccination, thus causing higher rates of reinfection [6, 46]. The S-protein in the SARS-CoV-2 genome carries many mutations that contribute to the successful dominance of some variants in the population [32]. For example, the N501Y mutation in the Alpha variant increases transmissibility by 70–80% compared to the original Wuhan-I strain. Similarly, the D614G mutation in the Delta variant enhances its transmissibility by 40–60% compared to the Alpha variant [32]. There are over 30 changes on the S-protein in the Omicron variant that lead to a faster spread, stronger immune escape, and enhanced binding to the ACE-2 receptor [32]. The S-protein has been the primary target for both vaccine and drug development.

Viral evolution may confer resistant mutations that pose a threat to antibody treatment. The S-protein on SARS-CoV-2 virus invades human cell by binding to the receptors. Monoclonal antibodies prevent this by attaching to the binding sites and forming a barrier [16]. Sotrovimab is one type of monoclonal antibody that neutralizes the SARS-CoV-2 virus and is available under emergency use to treat severely ill patients [34]. DR reduces treatment efficacy and increases the chance of failure. One study in Australia during a Delta-variant outbreak in 2021 found 4 out of 100 patients receiving sotrovimab acquired resistant mutations following treatment [34]. Likewise, DR also reduces the efficacy of direct-acting antivirals (DAAs). DAA treatment works by interfering with viral replication, either via suppressing or preventing replication, which reduces the risk of severe disease outcomes and may help a more rapid recovery [15]. There are two types of DAAs: the biological agent and the small molecule. The latter includes molnupiravir by Merck and Paxlovid by Pfizer [15]. These drugs are most effective during the early stages of SARS-CoV-2 infections with high viral replication [15].

Many drug development efforts for treatment of COVID-19 focus on interfering with SARS-CoV-2 proteases [38]. For example, Paxlovid quashes COVID-19 by targeting the main protease (Mpro) [22]. The active ingredient in Paxlovid that inhibits the production of Mpro is nirmatrelvir [38]. When a drug is used, disadvantaged mutants (e.g., those with high in-host fitness costs) can be favored and gain competitive advantage in replication [38]. If mutations conferring DR at the binding sites are already present at the time that treatment is initiated, they can reduce their fitness costs as a result of selection, potentially outcompeting drug-sensitive (DS) mutants and becoming dominant for transmission [38]. If mutations exist in proximity to the binding sites, the drug efficacy may be reduced with a lower binding energy [38]. Such mutations have already been found in some infected individuals [40]. Since many positions on viral proteases have mutated (e.g., P132H in the case of Omicron), Paxlovid as the current most effective and widely used oral antiviral treatment can become less effective in fighting the disease [22, 38, 40]. The rise in prescriptions for Paxlovid will inevitably exert an increasing selective

pressure on the virus over time and could facilitate the population growth and spread of resistant mutants [38, 40].

Given the above considerations, we sought to investigate the transient and long-term dynamics of DR based on the evolution of resistant mutants prior to the start of treatment. Experimental studies suggested that such mutations may incur at a fitness cost, impairing their ability to grow and propagate efficiently in the absence of treatment [24]. Our model accounts for this fitness cost and evaluates scenarios of treatment coverage in which DR becomes a population phenomenon. We will first quantify the contribution of DR infection to the total incidence under the scenario of long-term protection from primary infection. We then consider the possibility of reinfection with the same variant due to short-term immune protection. These scenarios will be investigated by varying the rate of resistant mutants emerging prior to the start of treatment and the rate of DR developing with a relative transmission probability reflecting the fitness cost during the treatment. This analysis will provide important insights for drug therapy as emerging mutations that exhibit DR undermine the efficacy of existing treatments, especially those with higher infectivity and fatality rates [20].

7.2 Methods

7.2.1 Model Structure

To construct the model, we classified individuals by the following epidemiological statuses: susceptible (S), infected with the DS variant (but not yet infectious) (E_S), asymptomatic (and infectious) (A_S), presymptomatic (and infectious) (P_S), symptomatic (and infectious) (I_S), and recovered (R) (Fig. 7.1). A proportion (p) of those who are infected will be asymptomatic, and the remaining will proceed to presymptomatic stage, ultimately developing symptoms. We assume that a proportion of symptomatic cases receive treatment (I_{ST}).

The emergence of DR in our model relies on the existence of resistant mutants as a result of viral replication prior to the start of treatment. We assume that, at the start of treatment, a proportion (r) of symptomatic cases will have the viral population that includes DR mutants if they were initially infected with the DS variant. We considered a probability α that symptomatic individuals harboring DR mutants will develop DR under the selection pressure of treatment (I_R) and therefore contribute to the spread of DR infection. Those who are not treated, whether harboring only DS viruses (I_S) or both DS and DR mutants (I_{SR}), can only spread DS infection [1, 42]. We assume that symptomatic cases who start treatment without having DR mutants (I_{ST}) will not develop DR. Treatment does not necessarily lead to the emergence of DR. Thus, with a probability of $1 - \alpha$, symptomatic cases with both DR and DS viral populations (I_{SRT}) will not develop DR under selection pressure of treatment. Once DR is emerged in treated cases, it can be spread with similar transmission dynamics

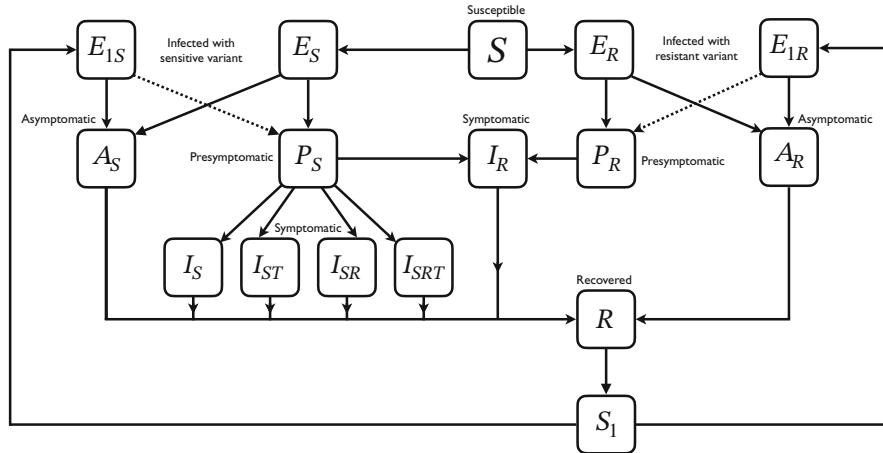


Fig. 7.1 Schematic diagram of disease transmission dynamics with treatment and emergence of DR

to the DS, including classes of infected (but not yet infectious) (E_R), asymptomatic (and infectious) (A_R), presymptomatic (and infectious) (P_R), and symptomatic (and infectious) (I_R) individuals (Fig. 7.1).

7.2.2 Model Equations

Disease transmission occurs as a result of contacts between susceptible and infectious individuals at a rate (β_S) for the DS infection. We assigned a parameter (c) of relative transmissibility for the DR infection to account for the fitness cost associated with the emergence of DR. Transition rates between different epidemiological classes are summarized in Table 7.1.

To investigate the parameter space in our model, we considered scenarios of short- and long-term protection after recovery from the DR or DS infection. In the case of short-term protection against reinfection, we introduced a parameter ϕ to represent the rate of loss of immunity. When $\phi > 0$, recovered individuals whose immunity against infection has waned transit to a susceptible state (S_1) and may experience reinfection (Fig. 7.1). Several studies have shown a decreased risk of reinfection by 80–93% for at least 6–9 months following recovery from primary infection [10, 13, 14, 31, 41], while other studies have found no decline in protection within 12 months post primary infection [9, 45]. Therefore, we assumed that the risk of reinfection may be lower by introducing a parameter ζ , where $\zeta \leq 1$, to represent the reduction in susceptibility to reinfection.

The dynamics of disease transmission are then expressed through proportional mixing by the following system of differential equations:

Table 7.1 Description of model parameters, their estimated values, and sources

Symbol	Description	Value	Source
β_S	Transmission rate for DS infection	0.52668	Calculated for $R_0 = 2.38$ [21]
c	Relative transmissibility of DR infection	0–1	Varied
p	Probability of becoming asymptomatic in primary infection	0.351	[36]
ζ	Reduction in susceptibility to reinfection	80–93%	[10, 13, 14, 41]
p_1	Probability of becoming asymptomatic in reinfection	0.7	Assumed
$1/\sigma$	Latent period	2.2	[21, 27]
$1/\theta$	Presymptomatic period	2.3	[26]
$1/\gamma_A$	Infectious period of asymptomatic	5	[21, 26]
$1/\gamma$	Infectious period of symptomatic cases	3.2	[26]
$1/\delta$	Infectious period for DS infection under treatment	2.2	Assumed
η	Probability of receiving treatment	0–1	Varied
r	Probability of DR mutants emerging	0–0.1	Assumed
α	Probability of developing DR during treatment	0–0.1	Assumed
κ	Transmissibility of asymptomatic infection relative to presymptomatic	0.26	[26, 37]
ϵ	Transmissibility of symptomatic infection relative to presymptomatic	0.89	[26]
ν	Transmissibility of symptomatic infection under treatment relative to presymptomatic	0.2	Assumed
$1/\phi$	Duration of immune protection against reinfection	≥ 0	Varied

$$\begin{aligned}
S' &= -\beta_S S \Lambda_S - \beta_R S \Lambda_R, \\
E'_S &= \beta_S S \Lambda_S - \sigma E_S, \\
E'_{1S} &= \zeta \beta_S S_1 \Lambda_S - \sigma E_{1S}, \\
A'_S &= p \sigma E_S + p_1 \sigma E_{1S} - \gamma_A A_S \\
P'_S &= (1-p) \sigma E_S + (1-p_1) \sigma E_{1S} - \theta P_S, \\
I'_S &= (1-\eta)(1-r)\theta P_S - \gamma I_S \\
I'_{SR} &= (1-\eta)r\theta P_S - \gamma I_{SR} \\
I'_{ST} &= \eta(1-r)\theta P_S - \delta I_{ST} \\
I'_{SRT} &= (1-\alpha)\eta r \theta P_S - \delta I_{SRT} \\
E'_R &= \beta_R S \Lambda_R - \sigma E_R, \\
E'_{1R} &= \zeta \beta_R S_1 \Lambda_R - \sigma E_{1R},
\end{aligned}$$

$$\begin{aligned}
A'_R &= p\sigma E_R + p_1\sigma E_{1R} - \gamma_A A_R \\
P'_R &= (1-p)\sigma E_R + (1-p_1)\sigma E_{1R} - \theta P_R, \\
I'_R &= \theta P_R + \alpha\eta r\theta P_S - \gamma I_R \\
R' &= \gamma_A(A_S + A_R) + \gamma(I_S + I_{SR} + I_R) + \delta(I_{ST} + I_{SRT}) - \phi R \\
S'_1 &= \phi R - \zeta\beta_S S_1 \Lambda_S - \zeta\beta_R S_1 \Lambda_R,
\end{aligned}$$

where $\beta_R = c\beta_S$,

$$\begin{aligned}
\Lambda_S &= (\kappa A_S + P_S + \epsilon I_S + \nu\epsilon I_{ST} + \epsilon I_{SR} + \nu\epsilon I_{SRT})/N \\
\Lambda_R &= (\kappa A_R + P_R + \epsilon I_R)/N,
\end{aligned}$$

are the forces of infection for the DS and DR viruses, respectively, and N represents the total population size. Assuming a constant population size, we have omitted demographic factors of birth and death in this model.

7.2.3 Reproduction Number

To assign a value to the transmission rate β_S , we first calculated the basic reproduction number of the model in the absence of control measures, which is the treatment in our model. Without treatment, the model reduces to a system of equations for the dynamics of DS infection only, as expressed by:

$$\begin{aligned}
S' &= -\beta_S S \Lambda_S, \\
E'_S &= \beta_S S \Lambda_S - \sigma E_S, \\
E'_{1S} &= \zeta\beta_S S_1 \Lambda_S - \sigma E_{1S}, \\
A'_S &= p\sigma E_S + p_1\sigma E_{1S} - \gamma_A A_S \\
P'_S &= (1-p)\sigma E_S + (1-p_1)\sigma E_{1S} - \theta P_S, \\
I'_S &= (1-r)\theta P_S - \gamma I_S \\
I'_{SR} &= r\theta P_S - \gamma I_{SR} \\
R' &= \gamma_A A_S + \gamma(I_S + I_{SR}) - \phi R \\
S'_1 &= \phi R - \zeta\beta_S S_1 \Lambda_S,
\end{aligned}$$

with

$$\Lambda_S = (\kappa A_S + P_S + \epsilon I_S + \epsilon I_{SR})/N$$

Although the evolution of resistant mutants is assumed to be outcompeted by higher fitness of DS replication in the absence of treatment, we retain the two compartments of I_S and I_{SR} as separate classes for clarity. Using the next-generation matrix approach [44], the basic reproduction number is expressed by

$$R_0 = \beta_S \left(\frac{p\kappa}{\gamma_A} + \frac{1-p}{\theta} + \frac{(1-p)\epsilon}{\gamma} \right) \quad (7.1)$$

7.3 Results

To parameterize the model, we calculated the transmission rate β_S for the DS virus by setting $R_0 = 2.38$ [21] and fixing other parameters in their ranges as given in Table 7.1. We then determined the transmission rate β_R for the DR mutants by varying the relative transmissibility.

7.3.1 Baseline Scenario

When considering a long-term immunity conferred by primary infection ($\phi = 0$), outbreak scenarios mimic the pattern of the classical SIR epidemic model with a single wave of infection. In this case, we evaluated the contribution of DR infections to the overall incidence of disease throughout the epidemic by varying the proportion of symptomatic cases that are treated.

Assuming a 60% relative transmissibility ($c = 0.6$) of DR, not surprisingly, higher treatment levels of symptomatic cases resulted in greater development of DR and, consequently, a larger contribution of DR to the overall incidence. Figure 7.2 shows that up to 2.5% of the overall incidence can be caused by DR infections when r and α vary in the range 0–10% and up to 75% of symptomatic cases are treated. As shown by contour curves in Fig. 7.2, probabilities of DR mutant emergence and DR development under treatment have a similar proportional effect on contribution of DR disease spread.

We further evaluated the optimal treatment level (η) which minimizes the overall attack rate (i.e., the proportion of the population infected throughout the epidemic) with varying probabilities of DR mutants arising and DR development during the treatment. For example, when the relative transmissibility of DR was 80%, Fig. 7.3 shows that for nonzero, but small r or α (e.g., $\leq 5 \times 10^{-4}$), the treatment of symptomatic cases can increase to high levels (i.e., almost full coverage) for minimizing the attack rate. However, as these probabilities increase, the optimal treatment level at which the attack rate is minimum reduces. We found that, because of low fitness costs of DR in this case, even a small increase in the treatment beyond its optimal level can lead to a relatively significant increase in the overall attack rate, especially for small r and/or α (Fig. 7.3).

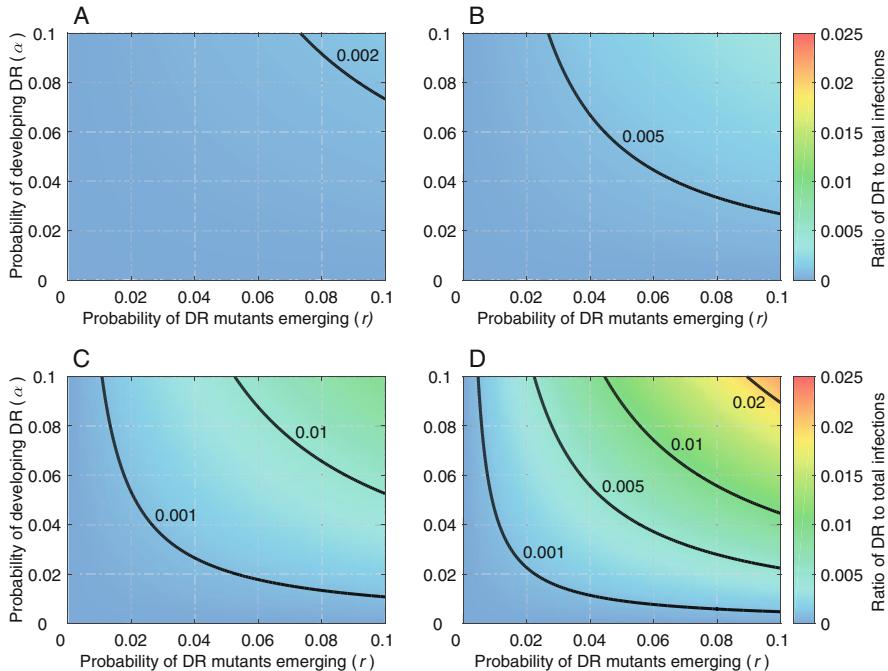


Fig. 7.2 Ratio of cumulative DR incidence to the total infections throughout the outbreaks as a function of the probability of DR mutants emerging (r) and the probability of developing DR during treatment (α). The reproduction number of DS virus was $R_0 = 2.38$, and the relative transmissibility of DR mutants was set to 60%. Assuming no waning immunity after recovery from primary infection, the treatment level (η) of symptomatic cases was set to (a) 10%, (b) 25%, (c) 50%, and (d) 75%

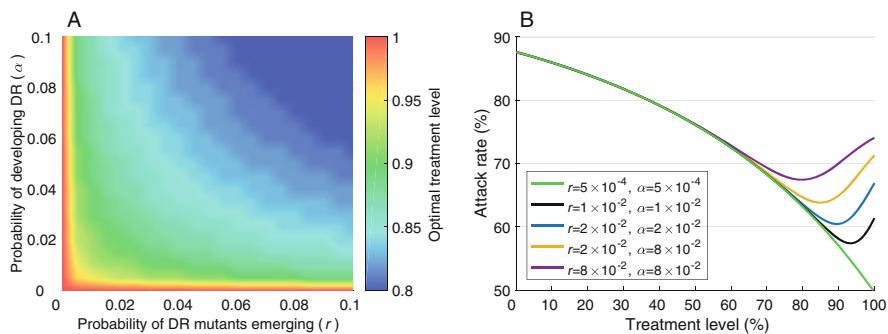


Fig. 7.3 (a) Optimal treatment level that minimizes the attack rate as a function of the probability of DR mutants emerging (r) and the probability of developing DR during treatment (α). The reproduction number of DS virus was $R_0 = 2.38$, and the relative transmissibility of DR mutants was set to 80%. (b) The overall attack rate with fixed r and α as a function of treatment level

7.3.2 Waning Immunity and Reinfection

We extended our analysis to consider the dynamics of DR infections over time when immunity acquired by prior infection is transient and reinfection can occur. We assumed an average duration of 180 days for protection against reinfection in recovered individuals and set $\zeta = 0.15$ for the reduction of susceptibility to reinfection. In this case, we observed the occurrence of more than one infection wave, with a trend of disease incidence that approaches an endemic state (Fig. 7.4).

For three sets of parameters associated with probabilities of DR mutants emerging and DR developing under treatment, we simulated the model to derive the proportion of daily incidence caused by DR infections (Fig. 7.4). Assuming a 60% relative transmissibility of DR, we found that the proportion of daily incidence caused by DR infections remained relatively low with a peak of less than 0.04% when 75% of symptomatic cases were treated. This proportion was less than 0.01% at the peak for lower treatment levels (i.e., $\leq 50\%$) even at the highest probabilities $r = 0.1$ and $\alpha = 0.1$ simulated here.

With a lower fitness cost of DR, i.e., a higher relative transmissibility ($c = 0.8$), the peak proportion of incidence caused by resistance can exceed 0.01% even if under 50% of symptomatic cases are treated. For example, if 25% of symptomatic cases are treated, the peak proportion exceeds and remains above 0.01% when $r = \alpha = 0.1$ (Fig. 7.4f). For 50% treatment level, this proportion increases to 2.5% within 3 years (Fig. 7.4g). When the treatment level increased to 75%, the DR

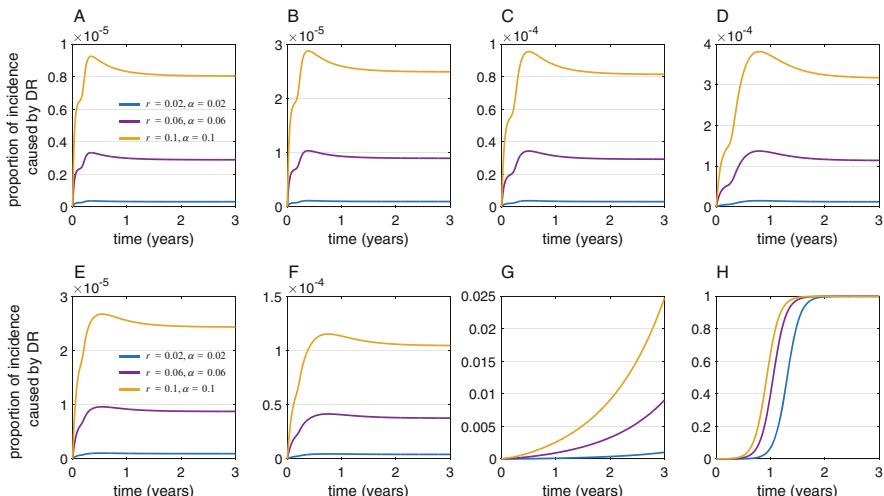


Fig. 7.4 Proportion of the total incidence caused by DR infections with waning immunity and reinfection over time. The reproduction number of DS virus was $R_0 = 2.38$, and the relative transmissibility of DR mutants was set to (a–d) 60% and (e–h) 80%. The probabilities of DR mutants emerging (r) and developing DR during treatment (α) were fixed as indicated in the legend, and the treatment level was set to (a, e) 10%, (b, f) 25%, (c, g) 50%, and (d, h) 75%

infections outcompeted DS infections, leading to the dominance of DR incidence over time. In this case, the proportion of incidence caused by DR infections approaches 100% within 2 years in a logistic-like curve (Fig. 7.4h). The probabilities r and α affected the rate of increase in DR incidence, with a shorter timeline to approach saturation as r and/or α increases.

7.4 Discussion

Emergence of DR remains an important consideration in the use of antiviral drugs for the treatment of SARS-CoV-2 infections. Our study aims to dissect the dynamics of DR based on a number of key parameters, including the probability of resistant mutants emerging prior to the start of treatment, the probability of developing DR infection during the treatment, and the transmissibility of DR relative to the DS virus. As we have illustrated by means of simulations, these parameters play a crucial role in determining the extent to which DR can contribute to the spread of infection and the treatment levels that can reduce the spread of disease without promoting the emergence and spread of DR. Our results indicate that, within a plausible range of parameters used here, the population-wide spread of drug-resistance is unlikely in the near term. Its long-term prevalence however depends on the fitness cost of DR mutants compared to the DS viruses and strategies for antiviral treatment of symptomatic cases. Our findings highlight the importance of monitoring the emergence, transmissibility, and further evolution of SARS-CoV-2 variants and subvariants to rapidly identify mutants that are resistant to antiviral treatment. Identification of such mutants can help utilize strategies (e.g., case isolation) in order to prevent the spread of DR if developed under treatment and reduce the potential for subsequent erosion of antiviral effectiveness.

Our model considers the evolution of DR mutants before the start of treatment as a requirement for developing DR during the treatment. Previous studies have presented ways to model SARS-CoV-2 mutations, such as cascade modeling based on the number and probability of mutations, Bayesian inference considering travel and variant competition, and spreading mutations based on viral RNA sequences [11, 19, 23]. However, there was an absence of extended traditional compartmental models that directly connect DR infections with mutations, treatment level, and fitness cost.

Several studies have reported the notable impact of mutated variants on transmission risk, immune escape, and the risk of reinfection [5, 6]. The rapid evolution of SARS-CoV-2 variants may generate DR mutants and can potentially lead to the rise of DR infections as the antiviral drugs are more widely prescribed. Encouragingly, however, there are a number of factors that may attenuate this potential. For example, the rise of immunity in the populations through vaccination or natural infection has dramatically reduced the severity of the disease, increasing the likelihood of asymptomatic or mild symptomatic (re)infections for which antiviral treatment is unlikely to be prescribed. Additionally, antiviral stewardship for treatment of high-

risk individuals and severely ill patients will limit the use of drugs and consequently the emergence of DR.

While our model encapsulates the natural history of the disease caused by SARS-CoV-2, the results must be interpreted within the context of study assumptions and limitations. Our model is parameterized based on the current knowledge and estimates. At present, there are no specific estimates pertaining to the parameters of DR mutant emergence and fitness cost; thus parameterization is based on previous knowledge and feasible ranges. We assumed that DR mutants emerging following infection with drug-sensitive mutants cannot be transmitted in the absence of selection pressure of treatment. Although this assumption is reasonable for mutants with a high fitness cost, compensatory mutations can arise to improve this fitness [12]. We also made several assumptions with respect to the effect of treatment. For instance, we assumed a shorter infectious period for symptomatic cases that are treated, with a lower transmissibility due to the effect of treatment that is impeding viral replication. Since COVID-19 drugs such as Paxlovid are relatively new, there is limited clinical data to assess their effect on the reduction of infectious period and infectiousness. Furthermore, the degree of drug selection pressure during treatment is undetermined. Our model does not explicitly include the wide spectrum of reality for SARS-CoV-2 infection, such as vaccination, application of other measures (e.g., isolation of cases), pre-existing immunity, demographics and geographic attributes, risk factors, or variability in disease-specific characteristics. However, the reproduction number accounts for the absence of these factors. Nevertheless, as proof of concept, our model highlights the importance of key parameters in the long-term dynamics of drug-resistance. Quantification of these parameters remains an important task for future studies.

Acknowledgments This work was partially supported by the NSERC-PHAC Emerging Infectious Disease Modeling initiative funded by the Mathematics for Public Health (MfPH) network. SM and JW were supported by the NSERC Discovery Grants. They also acknowledge the support of CIHR COVID-19 Rapid Response program.

References

1. Alexander, M.E., Bowman, C.S., Feng, Z., Gardam, M., Moghadas, S.M., Röst, G., Wu, J., Yan, P.: Emergence of drug resistance: implications for antiviral control of pandemic influenza. *Proc. R. Soc. B: Biol. Sci.* **274**(1619), 1675–1684 (2007)
2. Allué-Guardia, A., García, J.I., Torrelles, J.B.: Evolution of drug resistant mycobacterium tuberculosis strains and their adaptation to the human lung environment. *Front. Microbiol.* **12**, 612675 (2021)
3. Arino, J., Bowman, C.S., Moghadas, S.M.: Antiviral resistance during pandemic influenza: implications for stockpiling and drug use. *BMC Infect. Dis.* **9**(1), 1–12 (2009)
4. Center for Disease Control and Prevention: Covid-19 and antimicrobial resistance (2022). <https://www.cdc.gov/drugresistance/covid19.html>. Accessed 1 Oct 2022

5. Chadha, J., Khullar, L., Mittal, N.: Facing the wrath of enigmatic mutations: a review on the emergence of severe acute respiratory syndrome coronavirus 2 variants amid coronavirus disease-19 pandemic. *Environ. Microbiol.* **24**(6), 2615–2629 (2022)
6. Chakraborty, C., Bhattacharya, M., Sharma, A.R.: Present variants of concern and variants of interest of severe acute respiratory syndrome coronavirus 2: their significant mutations in s-glycoprotein, infectivity, re-infectivity, immune escape and vaccines activity. *Rev. Med. Virol.* **32**(2), e2270 (2022)
7. Davies, J., Davies, D.: Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**(3), 417–433 (2010)
8. Guillermot, D., Courvalin, P., French Working Party to Promote Research to Control Bacterial Resistance: better control of antibiotic resistance. *Clin. Infect. Dis.* **33**(4), 542–547 (2001)
9. Gundlapalli, A.V., Salerno, R.M., Brooks, J.T., Averhoff, F., Petersen, L.R., McDonald, L.C., Iademarco, M.F., CDC COVID-19 Response Carroll Yulia I MD, PhD Gorwitz Rachel PhD Lin-Gibson Sheng Wang Lili Pinto Ligia A: SARS-CoV-2 serologic assay needs for the next phase of the us covid-19 pandemic response. In: Open Forum Infectious Diseases, vol. 8, p. ofaa555. Oxford University Press, Oxford (2021)
10. Hall, V.J., Foulkes, S., Charlett, A., Atti, A., Monk, E.J., Simmons, R., Wellington, E., Cole, M.J., Saei, A., Ogutu, B., et al.: SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: a large, multicentre, prospective cohort study (siren). *Lancet* **397**(10283), 1459–1469 (2021)
11. Halley, J.M., Vokou, D., Pappas, G., Sainis, I.: SARS-CoV-2 mutational cascades and the risk of hyper-exponential growth. *Microbial Pathogen.* **161**, 105237 (2021)
12. Handel, A., Regoes, R.R., Antia, R.: The role of compensatory mutations in the emergence of drug resistance. *PLoS Comput. Biol.* **2**(10), e137 (2006)
13. Hansen, C.H., Michlmayr, D., Gubbels, S.M., Mølbak, K., Ethelberg, S.: Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study. *Lancet* **397**(10280), 1204–1212 (2021)
14. He, Z., Ren, L., Yang, J., Guo, L., Feng, L., Ma, C., Wang, X., Leng, Z., Tong, X., Zhou, W., et al.: Seroprevalence and humoral immune durability of anti-SARS-CoV-2 antibodies in Wuhan, China: a longitudinal, population-level, cross-sectional study. *Lancet* **397**(10279), 1075–1084 (2021)
15. Horby, P., Barclay, W., Hiscox, J., Chand, M., Breuer, J., Sherwood, E., Owen, A.: Nervtag: Antiviral drug resistance and the use of directly acting antiviral drugs (DAAS) for covid-19, 8 December 2021. GOV.UK (2021)
16. Jahanshahlu, L., Rezaei, N.: Monoclonal antibody as a potential anti-covid-19. *Biomed. Pharmacother.* **129**, 110337 (2020)
17. Klugman, K.P., Black, S.: Impact of existing vaccines in reducing antibiotic resistance: primary and secondary effects. *Proc. Natl. Acad. Sci.* **115**(51), 12896–12901 (2018)
18. Knippl, D., Röst, G., Moghadam, S.M.: Population dynamics of epidemic and endemic states of drug resistance emergence in infectious diseases. *PeerJ* **5**, e2817 (2017)
19. Lee, B., Sohail, M.S., Finney, E., Ahmed, S.F., Quadeer, A.A., McKay, M.R., Barton, J.P.: Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data (2022). medRxiv pp. 2021–12
20. Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., et al.: The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* **182**(5), 1284–1294 (2020)
21. Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., Shaman, J.: Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**(6490), 489–493 (2020)
22. MacPherson, K.: Key Antiviral Treatment for covid-19 Still Effective Despite Resistance Fears. Rutgers University (2022)

23. Maher, M.C., Bartha, I., Weaver, S., Di Iulio, J., Ferri, E., Soriaga, L., Lempp, F.A., Hie, B.L., Bryson, B., Berger, B., et al.: Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Sci. Transl. Med.* **14**(633), eabk3445 (2022)
24. Mari, A., Roloff, T., Stange, M., Søgaard, K.K., Asllanaj, E., Tauriello, G., Alexander, L.T., Schweitzer, M., Leuzinger, K., Gensch, A., et al.: Global genomic analysis of SARS-CoV-2 rna dependent rna polymerase evolution and antiviral drug resistance. *Microorganisms* **9**(5), 1094 (2021)
25. Moghadas, S.M., Bowman, C.S., Röst, G., Wu, J.: Population-wide emergence of antiviral resistance during pandemic influenza. *PLoS One* **3**(3), e1839 (2008)
26. Moghadas, S.M., Fitzpatrick, M.C., Sah, P., Pandey, A., Shoukat, A., Singer, B.H., Galvani, A.P.: The implications of silent transmission for the control of covid-19 outbreaks. *Proc. Natl. Acad. Sci.* **117**(30), 17513–17515 (2020)
27. Moghadas, S.M., Fitzpatrick, M.C., Shoukat, A., Zhang, K., Galvani, A.P.: Simulated identification of silent covid-19 infections among children and estimated future infection rates with vaccination. *JAMA Netw. Open* **4**(4), e217097–e217097 (2021)
28. Murray, C.J., Ikuta, K.S., Sharara, F., Swetschinski, L., Aguilar, G.R., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., et al.: Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**(10325), 629–655 (2022)
29. Nande, A.A.: Mathematical modeling of drug resistance and the transmission of SARS-CoV-2. Ph.D. Thesis, Harvard University (2021)
30. Organization, W.H.: Antimicrobial resistance (2021). <https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance>. Accessed 1 Oct 2022
31. Pilz, S., Chakeri, A., Ioannidis, J.P., Richter, L., Theiler-Schwetz, V., Trummer, C., Krause, R., Allerberger, F.: SARS-CoV-2 re-infection risk in Austria. *Eur. J. Clin. Invest.* **51**(4), e13520 (2021)
32. Prü, B.M.: Variants of sars cov-2: mutations, transmissibility, virulence, drug resistance, and antibody/vaccine sensitivity. *Front. Biosci. Landmark* **27**(2), 65 (2022)
33. Rice, L.B.: Antimicrobial stewardship and antimicrobial resistance. *Med. Clin.* **102**(5), 805–818 (2018)
34. Rockett, R., Basile, K., Maddocks, S., Fong, W., Agius, J.E., Johnson-Mackinnon, J., Arnott, A., Chandra, S., Gall, M., Draper, J., et al.: Resistance mutations in SARS-CoV-2 delta variant after sotrovimab use. *N. Engl. J. Med.* **386**(15), 1477–1479 (2022)
35. Rodrigues, P., Gomes, M.G.M., Rebelo, C.: Drug resistance in tuberculosis—a reinfection model. *Theor. Popul. Biol.* **71**(2), 196–212 (2007)
36. Sah, P., Fitzpatrick, M.C., Zimmer, C.F., Abdollahi, E., Juden-Kelly, L., Moghadas, S.M., Singer, B.H., Galvani, A.P.: Asymptomatic SARS-CoV-2 infection: a systematic review and meta-analysis. *Proc. Natl. Acad. Sci.* **118**(34), e2109229118 (2021)
37. Sayampanathan, A.A., Heng, C.S., Pin, P.H., Pang, J., Leong, T.Y., Lee, V.J.: Infectivity of asymptomatic versus symptomatic covid-19. *Lancet* **397**(10269), 93–94 (2021)
38. Sedova, M., Jaroszewski, L., Iyer, M., Godzik, A.: Monitoring for SARS-CoV-2 drug resistance mutations in broad viral populations (2022). bioRxiv
39. Sender, R., Bar-On, Y.M., Gleizer, S., Bernshtain, B., Flamholz, A., Phillips, R., Milo, R.: The total number and mass of SARS-CoV-2 virions. *Proc. Natl. Acad. Sci.* **118**(25), e2024815118 (2021)
40. Service, R.F.: Bad news for paxlovid? Coronavirus can find multiple ways to evade covid-19 drug (2022). <https://www.science.org/content/article/bad-news-paxlovid-coronavirus-can-find-multiple-ways-evade-covid-19-drug>. Accessed 1 Oct 2022
41. Sheehan, M.M., Reddy, A.J., Rothberg, M.B.: Reinfestation rates among patients who previously tested positive for coronavirus disease 2019: a retrospective cohort study. *Clin. Infect. Dis.* **73**(10), 1882–1886 (2021)
42. Strasfeld, L., Chou, S.: Antiviral drug resistance: mechanisms and clinical implications. *Infect. Dis. Clin.* **24**(3), 809–833 (2010)
43. Uchil, R.R., Kohli, G.S., KateKhaye, V.M., Swami, O.C.: Strategies to combat antimicrobial resistance. *J. Clin. Diagn. Res.* **8**(7), ME01 (2014)

44. Van den Driessche, P., Watmough, J.: Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **180**(1-2), 29–48 (2002)
45. Vitale, J., Mumoli, N., Clerici, P., De Paschale, M., Evangelista, I., Cei, M., Mazzone, A.: Assessment of SARS-cov-2 reinfection 1 year after primary infection in a population in Lombardy, Italy. *JAMA Internal Med.* **181**(10), 1407–1408 (2021)
46. Wang, R., Chen, J., Gao, K., Wei, G.W.: Vaccine-escape and fast-growing mutations in the united kingdom, the united states, Singapore, Spain, India, and other covid-19-devastated countries. *Genomics* **113**(4), 2158–2170 (2021)

Chapter 8

A Categorical Framework for Modeling with Stock and Flow Diagrams



John C. Baez, Xiaoyan Li, Sophie Libkind, Nathaniel D. Osgood, and Eric Redekopp

8.1 Introduction

Mathematical modeling of infectious disease at scale is important, but challenging. There are many benefits to the modeling process extending from taking diagrams as mathematical formalisms in their own right with the help of category theory. Stock and flow diagrams are widely used in infectious disease modeling, so we illustrate this point using these. However, rather than focusing on the underlying mathematics, we informally use communicable disease examples created with our software, called StockFlow.jl [32], to explain the benefits of the categorical framework. Readers interested in the mathematical details may refer to our earlier paper [7].

Many compartmental modelers regard diagrams offering a visual characterization of structure—e.g., susceptible, infective, and recovered stocks and transitions between them—as broadly accessible but informal steps towards a mathematically rigorous formulation in terms of ordinary differential equations (ODEs). However, ODEs are typically opaque to non-modelers—including the interdisciplinary members of the teams that typically are required for impactful models. By contrast, the system dynamics modeling tradition places a premium on engagement with stakeholders [15] and offers a modeling approach centered around diagrams. This approach commonly proceeds in a manner that depicts model structure using

J. C. Baez

Department of Mathematics, University of California, Riverside, CA, USA

e-mail: baez@math.ucr.edu

X. Li (✉) · N. D. Osgood · E. Redekopp

Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

e-mail: xiaoyan.li@usask.ca; nathaniel.osgood@usask.ca; eric.redekopp@usask.ca

S. Libkind

Department of Mathematics, Stanford University, Palo Alto, CA, USA

e-mail: slibkind@stanford.edu

successively more detailed models. The process starts with a “causal loop diagram” illustrating causal connections and feedback loops (Fig. 8.6). It then often proceeds to a “system structure diagram,” which distinguishes stocks from flows but still lacks quantitative information. The next step is to construct a stock and flow diagram—or henceforth, “stock-flow diagram” (Fig. 8.1). This diagram is visually identical to the system structure diagram, but it also includes formulae, values for parameters, and initial values of stocks.

The stock-flow diagram is treated as the durable end result of this modeling process, since it uniquely specifies a system of first-order ODEs. System dynamics modeling typically then alternates between assessing scenario outcomes resulting from numerically integrating the ODEs, performing other analyses (e.g., identifying location or stability of equilibria), and elaborating the stock-flow diagram—such as by adding elements to it, often borrowed from other models, or “stratifying” it by breaking large stocks (compartments) into smaller ones that differ in some characteristics.

While each of the types of diagrams in the system dynamics tradition is recommended by visual accessibility, the development of models using the traditional approach suffers from a number of practical shortcomings.

1. **Monolithic models:** Stock-flow models are traditionally built up in a monolithic fashion, leading ultimately to a single large piece of code. In larger models, this inhibits independent simultaneous work by multiple modelers. Lack of model modularity further prevents effective reuse of particular model elements. If elements of other models are used, they are commonly copy-and-pasted into the developing model, with the source and destination then evolving independently. Such separation can lead to a proliferation of conceptually overlapping models in which a single conceptual change (e.g., addition of a new asymptomatic infective compartment) requires corresponding updates in several successive models.
2. **Curse of stratification dimensionality:** While stratification is a key tool for representing heterogeneity and multiple lines of progression in compartmental models, it commonly requires modifications across the breadth of a model—stocks, flows, derived quantities, and many parameters. When that stratification involves multiple dimensions of heterogeneity, it can lead to a proliferation of terms in the ODEs. For example, rendering a model that characterizes COVID-19 into a model that also characterizes influenza would require that each COVID-19 state be replicated for each stage in the natural history of influenza. Represented visually, this stratification leads to a multidimensional lattice, commonly with progression proceeding along several dimensions of the lattice. Because of the unwieldy character of the diagram, much of the structure of the model is obscured. Adding, removing, or otherwise changing dimensions of heterogeneity routinely leads to pervasive changes across the model.
3. **Privileging ODE semantics:** The structure of causal loop diagrams, system structure diagrams, and stock-flow diagrams characterizes general state and accumulations, transitions, and posited causal relations—including induced feedbacks—amongst variables. Nothing about such a characterization restricts

its meaning to ordinary differential equations; indeed, many other interpretations and uses of these diagrams are possible. However, existing software privileges an ODE interpretation for stock-flow diagrams, while sometimes allowing for secondary analyses in ad hoc way—for example, identifying causal loops associated with the model or verifying dimensional homogeneity in dimensionally annotated models. Conducting other sorts of analyses—such as computation of eigenvalue elasticities or loop gains, analysis as a stochastic transition system, or other methods such as particle filtering [4, 18, 25, 28], particle MCMC [3, 19, 24], or Kalman filtered [13, 27] systems—typically requires bespoke software for reading, representing, and analyzing stock-flow models.

4. **Divergence of model representations:** Although the evolution from causal loop diagrams to system structure diagrams to stock-flow models is one of successive elaboration and informational enrichment, existing representations treat these as entirely separate characterizations and fail to capture the logical relationships between them. Such fragmentation commonly induces inconsistent evolution. Indeed, in many projects, the evolution of stock-flow diagrams renders the earlier, more abstract formulations obsolete, and the focus henceforth rests on the stock-flow diagrams.

What is less widely appreciated is that beyond their visual transparency and capacity to be lent a clear ODE semantics, both stock-flow diagrams themselves and their more abstract cousins possess a precise mathematical structure—a corresponding grammar, as it were. This algebraic structure, called the “syntax” of stock-flow diagrams, can be characterized using the tools of a branch of mathematics called category theory [12, 17]. Formalizing the syntax of stock-flow diagrams lends precise meaning to the process of “composing” such models (building them out of smaller parts), stratifying them, and other operations. Explicitly characterizing the syntax in software also allows for diagrams to be represented, manipulated, composed, transformed, and flexibly analyzed in software that implements the underlying mathematics.

Formalizing the mathematics of diagram-based models using category theory and capturing it in software offers manifold benefits. This paper discusses and demonstrates just a few:

1. **Separation of syntax and semantics.** Category theory gives tools to separate the formal structure, or “syntax,” of diagram-based models from the uses to which they are put, or “semantics.” This separation permits great flexibility in applying different semantics to the same model. With appropriate software design, this decoupling can allow the same software to support a diverse array of analyses, which can be supplemented over time.
2. **Reuse of structure.** The approaches explored here provide a structured way to build complex diagrams by composing small reusable pieces. With software support, modeling frameworks can allow for saving models and retrieving them for reuse as parts of many different models. For example, a diagram representing contact tracing processes can be reused across diagrams addressing different pathogens.

3. Modular stratification. A categorical foundation further supports a structured way to build stratified dynamical systems out of modular, reusable, largely orthogonal pieces. In contrast to the global changes commonly required for a diagram and the curse of dimensionality that traditionally arises when stratifying a diagram, categorically founded stratification methods allow for crisply characterizing a stratified diagram as built from simpler diagrams, one for each heterogeneity or progression dimension.

The balance of the chapter is structured as follows. Section 8.2 describes the categorical formulation of stock-flow diagrams. Section 8.3 explains how the categorical approach allows for decoupling the syntax and semantics of such diagrams. Section 8.4 introduces the hallmark of the categorical approach: the ability to build larger stock-flow diagrams from smaller pieces by composition. Section 8.5 discusses another key application of the categorical approach: stratification. Section 8.6 introduces ModelCollab: a web-based graphical user interface that allows users to build and run stock-flow models using the category-theoretic ideas we have introduced without requiring expertise in the mathematics. Finally, in Sect. 8.7 we close with some reflections on the significance and evolution of the methods described here. The code, for example, in this paper can be found in the Appendix.¹

8.2 The Syntax of Stock-Flow Diagrams

In this section, we illustrate the categorical method of representing stock-flow diagrams, which we can think of as characterizing the syntax of such diagrams. Section 3 of our previous work [7] introduced the mathematics underlying stock-flow diagrams. In this paper, we show how to encode an instance of a familiar stock-flow diagram—namely, the Susceptible-Exposure-Infectious-Recovered or “SEIR” model with an open population [2]—using the categorical method.

Figure 8.1 shows the stock-flow diagram for the SEIR model. The blue boxes labeled S, E, I, and R are “stocks.” The letter N is a “sum variable,” and there are blue “links” to it from the stocks on which it depends. The orange arrows are “flows.” Note that some flows go between stocks, while others enter a stock from outside the model, or leave a stock to go outside the model. The latter two cases are indicated with small “clouds.”

There are one or more blue links to each flow from the stocks and/or sum variables on which it depends. For each flow there is also a “flow variable” drawn in purple, which indicates the rate of that flow. Each flow variable is determined by some function of the stocks and/or sum variables that are connected to that flow variable by blue links. These functions are defined below the diagram in Fig. 8.1.

¹ The code can also be found in the GitHub repository <https://github.com/Xiaoyan-Li/applicationStockFlowMFPH>.

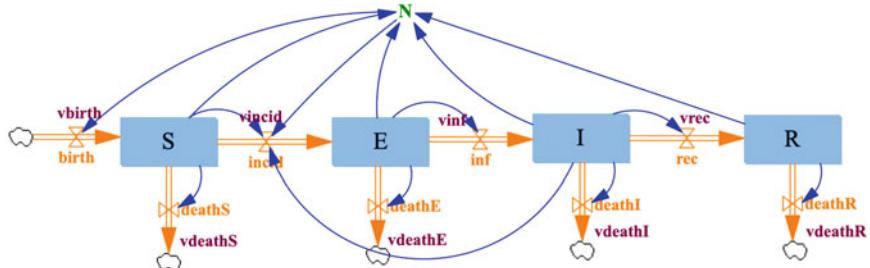


Fig. 8.1 The stock-flow diagram of the open population SEIR model

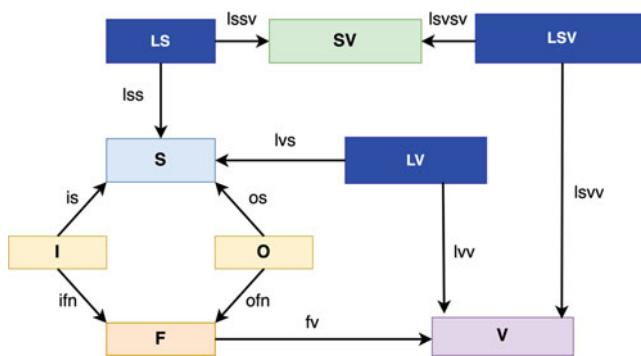


Fig. 8.2 The schema for stock-flow diagrams

For example, the flow describing infection, called inf , has a flow variable v_{inf} determined by the function f_{inf} .

The SEIR model is just one example of a stock-flow diagram. To formalize their general structure we use the so-called schema for stock-flow diagrams, shown in Fig. 8.2. It consists of boxes, called “objects,” and arrows between boxes, called “morphisms.” In an “instance” of this schema, we choose a set for each object and a function between such sets for each morphism. Other types of diagrams, such as causal loop diagrams, have their own schemas (see Sect. 8.3.2), and in fact there is a general theory of schemas [12, 30]. For now we consider only the schema for stock-flow diagrams. In this schema:

1. The objects S and F represent the stocks and flows, respectively.
2. The objects I and O represent the inflows and outflows. The morphisms $S \xleftarrow{\text{is}} F$ are used to describe which stocks are downstream of a given flow, while the morphisms $S \xleftarrow{\text{os}} O \xrightarrow{\text{ofn}} F$ are used to describe which stocks are upstream of a given flow. This structure allows for flows that go between two stocks, but also

flows that enter a stock from outside the model or leave a stock and go outside the model.

3. The object V represents “auxiliary variables,” sometimes termed “dynamic variables.” These are variables whose value is an instantaneous function of the current model state. The morphism $F \xrightarrow{fv} V$ indicates that the rate of each flow depends on one auxiliary variable. This relation is usually left implicit, not drawn, in the stock-flow diagram.
4. The object LV represents “variable links”: that is, links from stocks to auxiliary variables. The morphisms $S \xleftarrow{lvs} LV \xrightarrow{lvv} V$ indicate that any variable link goes from a stock to an auxiliary variable.
5. The objects SV and LS represent “sum variables” and “sum links.” Sum variables are a special type of auxiliary variable introduced in [7] to make composing stock-flow diagrams easier. A sum variable is simply the sum of the stocks linked to them by sum links. The arrows $S \xleftarrow{lss} LS \xrightarrow{lssv} SV$ indicate that any sum link goes from a stock to a sum variable.
6. The object LSV represents “sum variable links”: that is, links from sum variables to auxiliary variables. The arrows $SV \xleftarrow{lsvsv} LSV \xrightarrow{lsvv} V$ indicate that any sum variable link goes from a sum variable to an auxiliary variable.

An “instance” G of a schema assigns a finite set $G(X)$ to each object X of the schema and a function $G(a): G(X) \rightarrow G(Y)$ to each morphism $X \xrightarrow{a} Y$ of the schema. So, an instance of the schema for stock-flow diagrams consists of:

1. A finite set of stocks $G(S)$ and a finite set of flows $G(F)$
2. A finite set of inflows $G(I)$, a finite set of outflows $G(O)$, and functions

$$G(is): G(I) \rightarrow G(S), \quad G(ifn): G(I) \rightarrow G(F)$$

$$G(os): G(O) \rightarrow G(S), \quad G(ofn): G(O) \rightarrow G(F)$$

3. A finite set of auxiliary variables $G(V)$ and a function

$$G(fv): G(F) \rightarrow G(V)$$

4. A finite set of variable links $G(LV)$ and functions

$$G(lvs): G(LV) \rightarrow G(S), \quad G(lvv): G(LV) \rightarrow G(V)$$

5. A finite set of sum variables $G(SV)$, a finite set of sum links $G(LS)$, and functions

$$G(lss): G(LS) \rightarrow G(S), \quad G(lssv): G(LS) \rightarrow G(SV)$$

6. A finite set of sum variable links $G(\text{LSV})$ and functions

$$G(\text{lsvsv}): G(\text{LSV}) \rightarrow G(\text{SV}), \quad G(\text{lsvv}): G(\text{LSV}) \rightarrow G(\text{V})$$

A “stock-flow diagram” is a pair (G, ϕ) consisting of an instance G and, for each auxiliary variable v , a continuous function $\phi_v: \mathbb{R}^{G(\text{lvv})^{-1}(v)} \times \mathbb{R}^{G(\text{lsvv})^{-1}(v)} \rightarrow \mathbb{R}$. In Sect. 8.3.1 we explain how in the ODE semantics for stock-flow diagrams the function ϕ_v specifies how the value of the variable v depends on the stocks and sum variables that link to it.

Given an inflow $i \in G(\text{I})$ we say the stock $G(\text{is})(i)$ is “downstream” from the flow $G(\text{ifn})(i)$. Similarly, given an outflow $o \in G(\text{O})$ we say the stock $G(\text{os})(o)$ is “upstream” from the flow $G(\text{ofn})(o)$. In practice, we only want stock-flow diagrams where the functions $G(\text{ifn})$ and $G(\text{ofn})$ are injective. This constraint ensures that each flow has at most one downstream stock and at most one upstream stock. Also in practice we attach to each stock, flow, auxiliary variable, or sum variable an “attribute” which serves as its name. This naming relies on the theory of attributes developed by Patterson, Lynch, and Fairbanks [26].

When a stock-flow diagram is implemented in software, it is encoded as a categorical database [26]. Figure 8.3 shows the categorical database representing the stock-flow diagram for the SEIR model shown in Fig. 8.1. In this categorical database, each object X in the schema for stock-flow diagrams is represented by a database table. The row indices within this table consist of the elements of the set $G(X)$ associated to the object X . Thus, if the set $G(X)$ has n elements, then the table has n rows. In database parlance, each such row is associated with a “primary key” given in the first column of that table. For example, since there are four stocks in the SEIR model, the object S in the schema for stock-flow diagrams maps to the set $\{1, 2, 3, 4\}$, and the table for the object S has four rows numbered 1, 2, 3, 4. By contrast, the table for the object F has eight rows, reflecting the fact that there are eight flows in the SEIR model.

The table for any object X has one column for each morphism coming out of X , which describes the function associated to that morphism. For example, the table for the object LV has one column “lvs” describing the function $G(\text{lvs}): G(LV) \rightarrow G(S)$ mapping each variable link to a stock (as a “foreign key” giving the key of that stock in the “S” table), and one column “lvv” describing the function $G(\text{lvv}): G(LV) \rightarrow G(V)$ mapping each variable link to a variable (similarly specified by a foreign key). Besides this, the tables for S, F, V , and SV have an extra column giving names for the stocks, flows, and variables and sum variables, rather than foreign keys. Technically, these names are handled using the theory of attributes mentioned above.

This capacity to encode stock-flow diagrams in a mathematically precise and transparent fashion confers diverse benefits. To list a few, these include the capacity to compose such diagrams (see Sect. 8.4), to soundly transform such diagrams for optimization, and to parallelize them. But one of the most foundational benefits is the capacity to perform different types of analysis on such diagrams—that is, to

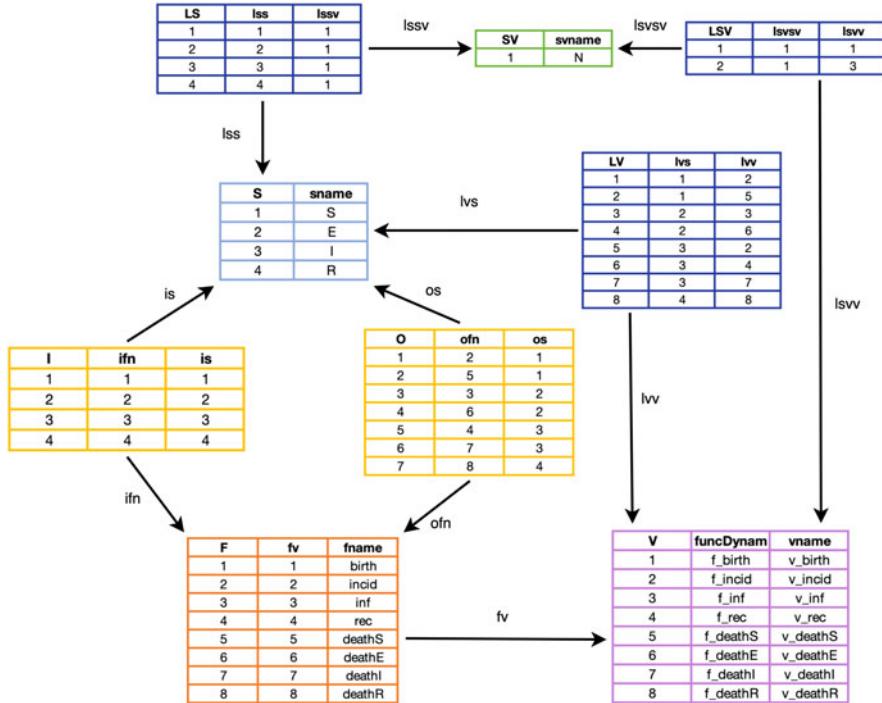


Fig. 8.3 The categorical database structure representing the stock-flow diagram associated with the SEIR model

interpret a given diagram using various choices of “semantics.” The next section discusses that benefit in greater detail.

8.3 The Semantics of Stock-Flow Diagrams

The capacity to interpret stock-flow diagrams in different ways is achieved by separating the syntax from the semantics of such diagrams. Matters of syntax concern the forms that such diagrams can take. In particular, the rules governing what counts as a legitimate stock-flow diagram—what can connect to what—are largely specified by the schema discussed in the previous section. But the syntax of stock-flow diagrams is distinct from the interpretation given to these diagrams. A given stock-flow diagram can be interpreted in different ways, each lending that diagram some meaning. Some of these interpretations involve dynamic simulations that solve equations specified by the diagram, while others describe static features of the diagram—for example, extracting algebraic equations specifying the equilibria of the system in terms of the parameters.

In this section, we introduce three choices of semantics for stock-flow diagrams that have been implemented in StockFlow.jl [7, 32]: ordinary differential equations and a pair of semantics that extract from a stock-flow diagram the associated causal loop diagram and system structure diagram.

8.3.1 ODEs (Ordinary Differential Equations)

A stock-flow diagram is traditionally used to represent a continuous-time, continuous-state dynamical system [31]: a system of ODEs that describes the evolution of each real-valued stock. While this addresses an important subclass of dynamical systems, this convention has needlessly restricted analysis potential and crimped the flexibility of supporting software by privileging a single interpretation of the syntax of the stock-flow diagrams. In this project, we mathematically decouple the choice of the stock-flow diagram (the syntax) from the choice of its interpretation (the semantics). As we shall see in this subsection, this approach readily supports the traditional interpretation of a stock-flow diagram in terms of ODEs. But as subsequent subsections illustrate, this interpretation is no longer required or even privileged.

The decoupling of syntax and semantics afforded by the categorical approach is achieved by the use of a structure-preserving map, called a “functor,” sending each stock-flow diagram to its interpretation, or meaning. The choice of different such functors allows for different interpretations.

“Functorial semantics”—the idea of treating semantics as a functor—goes back to Lawvere’s work [16] in the early 1960s. It has grown into a powerful method for specifying and analyzing the semantics of programming languages. By now, it has also been applied to many diagrammatic modeling languages, including Petri nets, electrical circuit diagrams and chemical reaction networks, and others [5, 6, 10]. Thus, the time is ripe for applying functorial semantics to stock-flow diagrams.

To do this, we need to define a *category* of stock-flow diagrams, which in essence means precisely defining not only these diagrams, as we have done above, but also maps between them. Using the methods of our previous paper [7] we can define such a category, called **StockFlow**, and also a functor

$$\text{StockFlow} \xrightarrow{v} \text{Dynam}$$

from this category to a category of ODEs and maps between those. Here, for simplicity, we simply explain how this functor sends any stock-flow diagram to an ODE.

For any stock $s \in G(S)$, the set of its inflows is $G(\text{is})^{-1}(s)$. Thus, this stock is downstream from precisely the flows in $G(\text{ifn})(G(\text{is})^{-1}(s))$. Similarly, this stock is upstream from precisely the flows in $G(\text{ofn})(G(\text{os})^{-1}(s))$. We denote as ϕ_v the continuous function describing the value of each auxiliary variable $v \in G(V)$

as a function of the stocks and sum variables linked to it, so $\phi_v: \mathbb{R}^{G(\text{lvv})^{-1}(v)} \times \mathbb{R}^{G(\text{lsvv})^{-1}(v)} \rightarrow \mathbb{R}$. Then, the function describing the rate of the flow $f \in G(\mathbf{F})$ is $\phi_{(G(\text{fv})(f))}$. The ordinary differential equation of the stock s can then be defined as follows:

$$\dot{s} = \sum_{f \in G(\text{ifn})(G(\text{is})^{-1}(s))} \phi_{(G(\text{fv})(f))} - \sum_{f \in G(\text{ofn})(G(\text{os})^{-1}(s))} \phi_{(G(\text{fv})(f))}, \quad (8.1)$$

Furthermore the value of each sum auxiliary variable is the sum of the values of the stocks it links to. For example, the value of the sum variable N is defined as $N = S + E + I + R$ in the SEIR model.

We can easily understand the functor mapping from a stock-flow diagram to the ODEs by the slogan: *the time derivative of the value of each stock equals the sum of its inflows minus the sum of its outflows*.

For example, the SEIR stock-flow diagram (Fig. 8.1) gives the following set of ODEs:

$$\begin{aligned} \dot{S} &= \mu N - \beta \frac{I}{N} S - \delta S \\ \dot{E} &= \beta \frac{I}{N} S - \frac{E}{t_{\text{latent}}} - \delta E \\ \dot{I} &= \frac{E}{t_{\text{latent}}} - \frac{I}{t_{\text{recovery}}} - \delta I \\ \dot{R} &= \frac{I}{t_{\text{recovery}}} - \delta R \end{aligned} \quad (8.2)$$

The solutions can be calculated and plotted out directly using the StockFlow.jl software, as in Fig. 8.4.

It is notable that the ODE semantics mapping is more general than illustrated above: we can also map the syntax of “open” stock-flow diagrams (to be discussed in Sect. 8.4) to “open” dynamical systems. This supports composition of models.

8.3.2 Causal Loop Diagrams

In this section, we define a semantics for stock-flow diagrams that maps any such diagram into a particularly simple kind of causal loop diagram. In system dynamics, a “causal loop diagram” is a graph where each node represents a variable and each edge represents a way in which one variable can directly influence another [31]. The edges are typically labeled with \pm signs called “polarities” that indicate whether an increase in the source variable tends to increase or decrease the target variable,

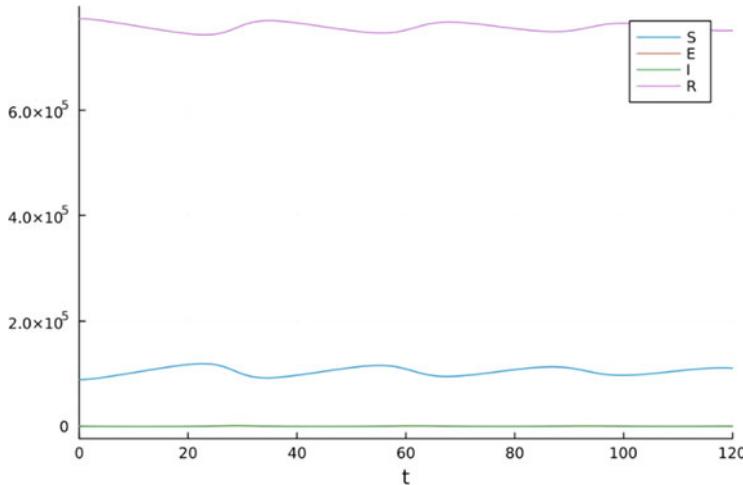
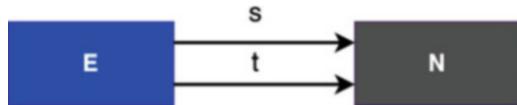


Fig. 8.4 An example solution of the ODEs of the SEIR stock-flow diagram

Fig. 8.5 The schema for causal loop diagrams



ceteris paribus. From these we can compute signs for loops of edges, which describe positive or negative feedback loops.

Here we consider a simplified variant of causal loop diagram which is simply a graph: the polarities for edges are not included. The problem of interpreting a stock-flow diagram as a fully annotated causal loop diagram is left for future work.

Figure 8.5 depicts the schema for causal loop diagrams. It is just the schema for what category theorists call “graphs” (i.e., directed multigraphs allowing self-loops). There is an object E for edges, an object N for nodes, and morphisms $s, t : E \rightarrow N$ sending each edge to its source and target node, respectively. Thus, an instance CL of this schema is simply a finite set $CL(N)$ of nodes, a finite set $CL(E)$ of edges, and maps $CL(s), CL(t) : CL(E) \rightarrow CL(N)$ sending each edge to its source and target node.

There is in fact a way to translate any stock-flow diagram (G, ϕ) into a causal loop diagram CL . It works as follows:

1. The set of nodes $CL(N)$ is the disjoint union of the set of stocks, the set of sum variables, and the set of auxiliary variables (which includes such a variable for each flow). Explicitly,

$$CL(N) := G(S) \sqcup G(SV) \sqcup G(V).$$

2. The set of edges $\text{CL}(E)$ is given by

$$\text{CL}(E) := \mathbf{G}(\text{LV}) \sqcup \mathbf{G}(\text{LS}) \sqcup \mathbf{G}(\text{LSV}) \sqcup \mathbf{G}(\text{I}) \sqcup \mathbf{G}(\text{O}).$$

3. Each edge coming from a variable link $\ell \in \mathbf{G}(\text{LV})$ has source $\mathbf{G}(\text{lvs})(\ell)$ and target $\mathbf{G}(\text{lvv})(\ell)$. The source and target of edges coming from sum links and sum variable links are defined similarly. Each edge coming from an inflow $i \in \mathbf{G}(\text{I})$ has the auxiliary variable $\mathbf{G}(\text{fv})\mathbf{G}(\text{ifn})(i)$ as its source and the stock $G(\text{is})(i)$ as its target. Note that the auxiliary variable $\mathbf{G}(\text{fv})\mathbf{G}(\text{ifn})(i)$ represents the flow for which i is the inflow. Similarly, each edge coming from an outflow $o \in \mathbf{G}(\text{O})$ has the stock $G(\text{os})(o)$ as its source and the auxiliary variable $\mathbf{G}(\text{fv})\mathbf{G}(\text{ofn})(o)$ as its target.

Using this procedure, the SEIR stock-flow diagram in Fig. 8.1 gives rise to the causal loop diagram in Fig. 8.6. Here, following tradition, we have labeled some of the nodes by flows rather than their corresponding auxiliary variables, e.g., “birth” rather than “vbirth.” Since the map $\mathbf{G}(\text{fv}): \mathbf{G}(\text{F}) \rightarrow \mathbf{G}(\text{V})$ is usually one-to-one, this does not create ambiguities in practice.

Just as with stock-flow diagrams, we use the data structure of a categorical database to encode causal loop diagrams in the software. Figure 8.7 shows an example: the categorical database for the SEIR causal loop diagram.

8.3.3 System Structure Diagrams

In system dynamics [31], a system structure diagram is a *purely qualitative* version of a stock-flow diagram. It lacks the quantitative information provided by the functions ϕ_v that describe how each auxiliary variable v depends on the quantities linked to it. While many uses of system structure diagrams further annotate links in such diagrams with polarities, we reserve for future work the problem of automatic derivation of such polarities. For now, we therefore define a “system structure

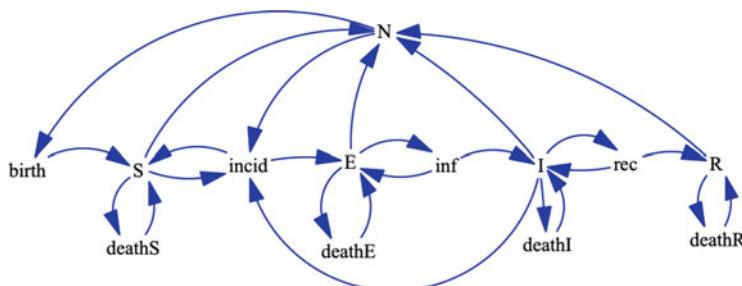


Fig. 8.6 The causal loop diagram of the SEIR model

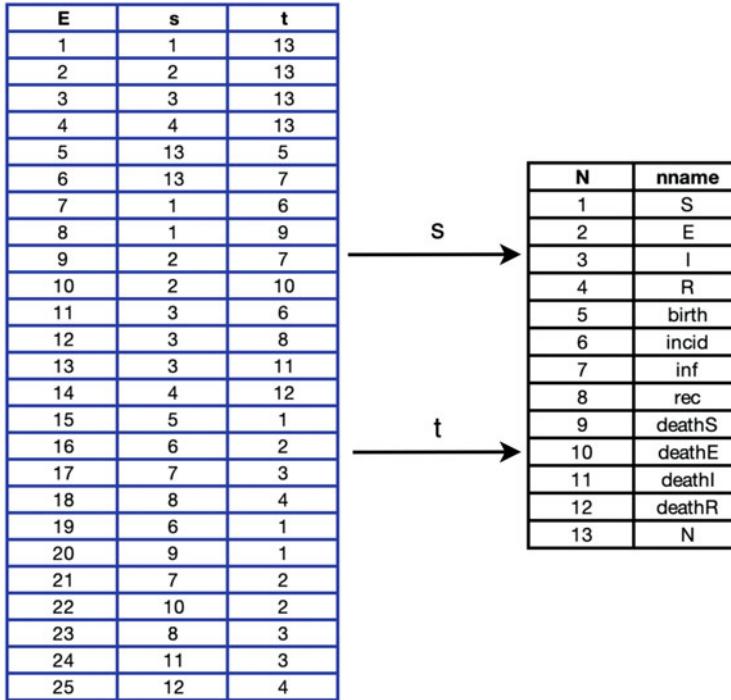


Fig. 8.7 The categorical database structure representing the SEIR causal loop diagram

diagram” to be simply an instance G of the schema for stock-flow diagrams, as explained in Sect. 8.2.

There is a semantics for stock-flow diagrams that maps them to system structure diagrams: it simply maps any stock-flow diagram (G, ϕ) to the system structure diagram G . In fact one can construct a category **StockFlow** of stock-flow diagrams, a category **SystemStructure** of system structure diagrams, and a functor

$$\text{StockFlow} \xrightarrow{A} \text{SystemStructure}$$

that maps any stock-flow diagram (G, ϕ) to the system structure diagram G .

Likewise, the semantics described in Sect. 8.3.2 gives a functor

$$\text{StockFlow} \xrightarrow{B} \text{CausalLoop}$$

from the category of stock-flow diagrams to a suitable category **CausalLoop**. But note that the causal loop diagram associated to a stock-flow diagram (G, ϕ) depends only on G : the functions ϕ_v play no role here, though they would for the more elaborate and more commonly used causal loop diagrams with edges labeled by

polarities. Thus, our causal loop and system structure semantics for stock-flow diagrams fit into a so-called commutative diagram of functors between categories:

$$\begin{array}{ccc} \text{StockFlow} & \xrightarrow{B} & \text{CausalLoop} \\ & \searrow A & \nearrow C \\ & \text{SystemStructure} & \end{array}$$

In essence, this diagram says that to turn a stock-flow diagram (G, ϕ) into a causal loop diagram in the manner explained in Sect. 8.3.2, we can first extract the system structure diagram G and then turn that into a causal loop diagram.

This commutative diagram illustrates a general fact: rather than the various semantics for a given form of syntax being separate from each other, like isolated walled gardens, they are often related in fruitful ways. Category theory lets us formalize these relationships, and category-based software lets us apply them in practical ways.

8.4 Composing Open Stock-Flow Diagrams

We can build larger stock-flow diagrams by gluing together smaller ones. To achieve this goal, we need “open” stock-flow diagrams. An example is shown in Fig. 8.8. It

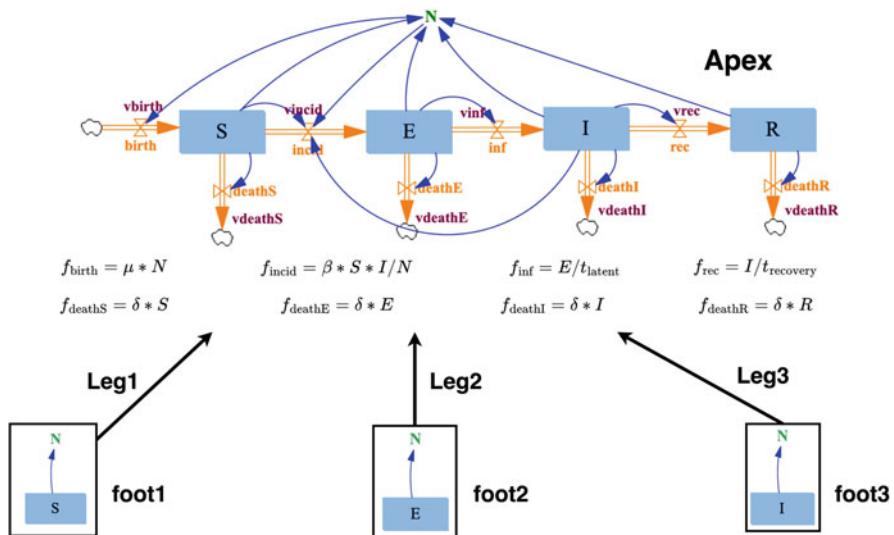


Fig. 8.8 An example of an open stock-flow diagram

consists of a stock-flow diagram together with some extra data describing interfaces at which we can compose this diagram to other stock-flow diagrams.

More precisely, an open stock-flow diagram consists of an “apex” together with a finite collection of “feet” and “legs.” The “apex” is any stock-flow diagram (G, ϕ) . Each “foot” is also a stock-flow diagram, and it comes with a “leg,” which is a map of stock-flow diagrams from the foot to (G, ϕ) . However, we require that the feet are stock-flow diagrams of a restricted sort: they can contain only stocks, sum variables, and sum links. We enforce this restriction because we want to glue together stock-flow diagrams only by identifying certain stocks in one diagram with stocks in another diagram; however, doing this properly may require identifying certain sum variables and sum links as well.

We call these restricted stock-flow diagrams “interfaces.” Just as there is a schema for stock-flow diagrams, there is a schema for interfaces, shown in Fig. 8.9. An interface, say X , is precisely an instance of this schema. It thus consists of:

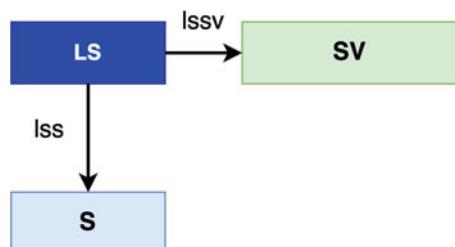
1. A finite set $X(S)$ of stocks
2. A finite set of sum variables $X(SV)$, a finite set of sum links $X(LS)$, and functions

$$X(\text{lss}) : X(LS) \rightarrow X(S), \quad X(\text{lssv}) : X(LS) \rightarrow X(SV)$$

In our previous paper [7] we explain how to compose open stock-flow diagrams. Briefly, we use the mathematics of “decorated cospans” [9]—a general category-theoretic framework for the composition of open systems. However, to implement this in code in StockFlow.jl we also take advantage of a closely related framework, “structured cospans” [5, 6]. The reason is that structured cospans have already been systematically implemented in the Catlab.jl package [26], which serves as the foundation underlying our implementation of StockFlow.jl. StockFlow.jl is one of the newer members of the AlgebraicJulia ecosystem [1], which provides computational support for applied category theory via Catlab.jl.

Two other twists are also worth noting, if only for cognoscenti. Firstly, in StockFlow.jl we actually use decorated or structured “multicospans” [21, 29] instead of cospans. The difference is that multicospans can have multiple legs and feet, as shown, for example, in Fig. 8.8, while cospans have exactly two. Secondly,

Fig. 8.9 The schema for interfaces



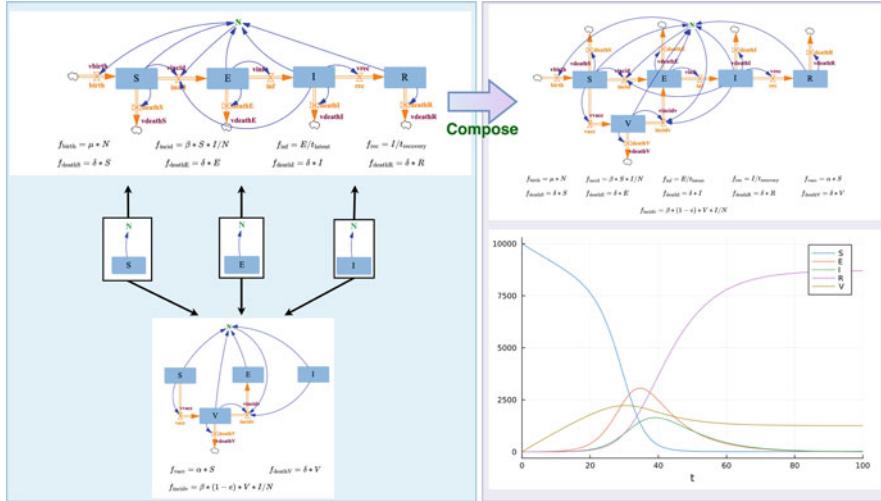


Fig. 8.10 Composing SEIR and SVEI models to obtain an SEIRV model

StockFlow.jl applies the flexible graphical syntax of undirected wiring diagrams [11] to compose these multicospans.

However, users do not need to understand these technicalities to use our software. The key ideas can be understood from an example. Figure 8.8 shows an open stock-flow diagram with three legs and feet: this is an SEIR model with three interfaces containing the stocks S, E, and I, respectively. Figure 8.10 shows an example of composition in which this open stock-flow diagram is glued to another open stock-flow diagram, an SVEI model, along all three interfaces. The result of composition is yet another open stock-flow diagram, but in this particular case there are no interfaces left, so it amounts to an ordinary stock-flow diagram. The composite diagram describes an SEIRV model. Like any other diagram, diagrams that result from such composition can be mapped to alternative semantic domains. For example, the bottom right plot in Fig. 8.10 then shows a solution of the system of ODEs to which the composite stock-flow diagram is mapped using ODE semantics.

8.5 Stratifying Typed System Structure Diagrams

Just as we introduced *open* stock-flow diagrams to permit *composition* of models, we now introduce *typed* diagrams to support *stratification* of models. Recall from Sect. 8.1 that “stratifying” a model involves breaking stocks into smaller stocks that differ in some characteristics. For example, we might take the simple SEIR model shown in Fig. 8.1 and subdivide each stock into two groups of different sexes, or three groups of different ages, or both. This subdivision is a common and important procedure for refining models in epidemiology.

However, stratification also requires introducing new flows, new auxiliary variables, new links, and so on. For example, consider the flow “inf” from the “Exposed” stock to the “Infectious” stock in the SEIR model. If we stratify this model by breaking each stock into two sexes, we also need to replace this flow with two separate flows, one for each sex. Furthermore, since the rate of this flow is given by an auxiliary variable, we must replace that variable by two separate variables if we wish to allow the two sexes to have different rates of infection—which indeed is the whole point of stratifying the model in this way.

The challenge is to carry out all these steps in a mathematically well-defined way that can be cleanly implemented in software. Libkind et al. recently did this for a related diagrammatic modeling language that has also been implemented in the AlgebraicJulia ecosystem: Petri nets [20]. The key was to introduce “typed” Petri nets and use pullbacks, a standard construction in category theory [17]. However, their approach to stratification has the potential to be generalized to many other diagrammatic modeling languages. Here we adapt their approach to stock-flow diagrams.

To begin, it is important to realize that when we stratify a model described by a stock-flow diagram (G, ϕ) , we do *not* expect that all the functions ϕ_v associated to auxiliary variables v can be copied over from the original model in an automatic way. Indeed, the point of stratification is to let these functions depend on the “stratum,” e.g., the sex, the age group, and so on. Thus, in the approach we take here, we first stratify not the whole stock-flow diagram (G, ϕ) but only its underlying system structure diagram G (as defined in Sect. 8.3.3). To promote the stratified system structure diagram to a stock-flow diagram, the user must then choose functions for the auxiliary variables. In future work, we can enable an approach where most, but not all, of the original functions are automatically reused.

How do we stratify system structure diagrams? A first naive thought would be to take a “product” of two system structure diagrams:

1. The original system structure diagram that we wish to stratify. We call this the “aggregate model” and denote it as $S_{\text{aggregate}}$.
2. A system structure diagram describing the strata (e.g., age groups or sexes) that we wish to use in stratifying the aggregate model. We call this the “strata model” and denote it as S_{strata} .

The product of these two, denoted $S_{\text{aggregate}} \times S_{\text{strata}}$, is a system structure diagram for which a stock is an ordered pair (x, y) consisting of a stock x in the aggregate model $S_{\text{aggregate}}$ and a stock y in the strata model S_{strata} . Similarly a flow in the product is an ordered pair of flows, one from each model—and so on for inflows, outflows, auxiliary variables, and all the other objects in the schema for stock-flow diagrams.

In fact, category theory has a general notion of “product” [17] applicable to any category, and $S_{\text{aggregate}} \times S_{\text{strata}}$ is a product in the category **SystemStructure**. One consequence is that it comes with maps as follows:

$$\begin{array}{ccc} S_{\text{aggregate}} \times S_{\text{strata}} & \xrightarrow{p_1} & S_{\text{aggregate}} \\ p_2 \downarrow & & \\ S_{\text{strata}} & & \end{array}$$

Given any ordered pair (x, y) of stocks, flows, etc. in the product $S_{\text{aggregate}} \times S_{\text{strata}}$, the maps p_1 and p_2 pick out the components of this ordered pair:

$$p_1(x, y) = x, \quad p_2(x, y) = y.$$

Unfortunately, the product $S_{\text{aggregate}} \times S_{\text{strata}}$ often contains more stocks, flows, etc. than we really want. The solution is to keep only the ordered pairs (x, y) where x and y have the same “type.” To do this, we introduce a third stock-flow diagram S_{type} , called the “type system,” together with maps

$$\begin{array}{ccc} & S_{\text{aggregate}} & \\ & \downarrow t_{\text{aggregate}} & \\ S_{\text{strata}} & \xrightarrow{t_{\text{strata}}} & S_{\text{type}} \end{array}$$

We obtain the desired stratified model $S_{\text{stratified}}$ by taking a “pullback” in the category **StockFlow**. The pullback is a particular stock-flow diagram equipped with maps making the following square commute:

$$\begin{array}{ccc} S_{\text{stratified}} & \xrightarrow{p_1} & S_{\text{aggregate}} \\ p_2 \downarrow & & \downarrow t_{\text{aggregate}} \\ S_{\text{strata}} & \xrightarrow{t_{\text{strata}}} & S_{\text{type}} \end{array}$$

The pullback is defined so that a stock in $S_{\text{stratified}}$ is an ordered pair (x, y) consisting of a stock x in the aggregate model $S_{\text{aggregate}}$ and a stock y in the strata model S_{strata} that both map to the same stock in the type system S_{type} . In other words, the pair (x, y) satisfies

$$(t_{\text{aggregate}})_S(x) = (t_{\text{strata}})_S(y).$$

Similarly, a flow in $S_{\text{stratified}}$ is a pair consisting of a flow in the aggregate model and a flow in the strata model that both map to the same flow in the type system—and so

on for inflows, outflows, auxiliary variables, and all the other objects in the schema for stock-flow diagrams. As before, the maps p_1 and p_2 pick out the components of these ordered pairs:

$$p_1(x, y) = x, \quad p_2(x, y) = y.$$

All this and more follows from the general theory of pullbacks, which works in any category [17]. That is why this approach to stratification works so generally.

Before we turn to examples of stratification, let us briefly explain the maps between system structure diagrams that appear as arrows in the above diagram. Technically these maps are the *morphisms* in the category **SystemStructure**, and they play an important role in the theory. But what are these maps like?

Given system structure diagrams G and H , a map $\alpha: G \rightarrow H$ consists of functions sending all the stocks, flows, inflows, outflows, auxiliary variables, etc. for G to the corresponding items for H . We use $\alpha_S: G(S) \rightarrow H(S)$ to denote the function on stocks, $\alpha_F: G(F) \rightarrow H(F)$ to denote the function on flows, and so forth. But we require that these functions be structure-preserving. For example, if α_F maps a flow $f \in G(F)$ to a flow $f' \in H(F)$, then we require that α_F must map the upstream of f to the upstream of f' , and map the downstream of f to the downstream of f' . For details, see [7].

Figure 8.11 shows an example of such a map α from the SEIR system structure diagram (top) to the SIR system structure diagram (bottom). To show the map clearly, we have colored the components of the SEIR diagram according to the color of the component in the SIR diagram to which they are mapped. Notice that α maps

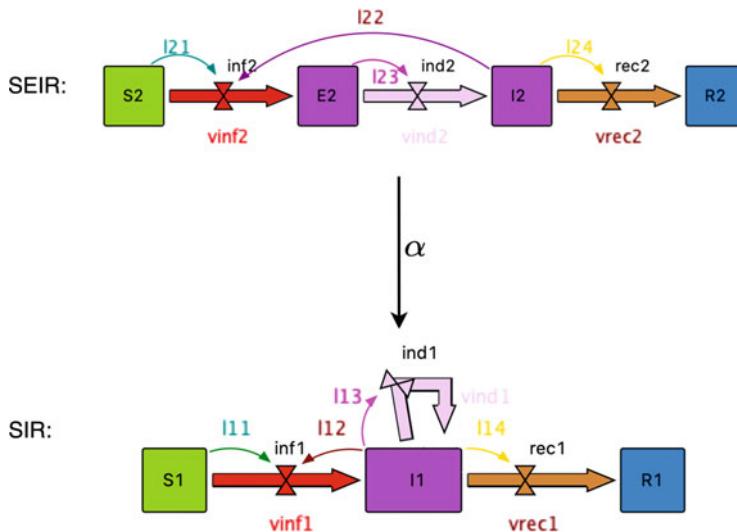


Fig. 8.11 An example of a map (morphism) from the SEIR system structure diagram to the SIR system structure diagram

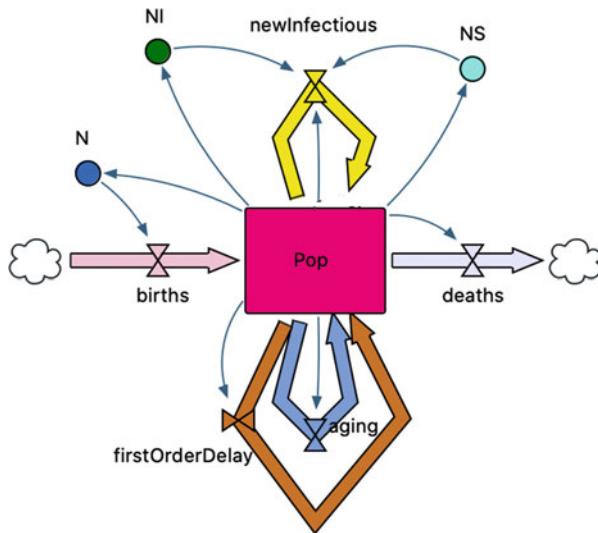


Fig. 8.12 A stock-flow diagram S_{type} serving as a type system for infectious disease models

the flow inf_2 of the SEIR diagram to the flow inf_1 of the SIR diagram. Thus, we require that α maps the upstream stock of inf_2 to the upstream stock of inf_1 , and the downstream stock of inf_2 to the downstream stock of inf_1 .

Now let us turn to examples of stratification. Figure 8.12 shows a system structure diagram S_{type} that can serve as a type system for stratified infectious disease models. This system structure diagram has just one stock, Pop , which represents all the kinds of populations. It has five flows, representing five different types of flows in the infectious disease models we wish to build:

1. The birth flow
2. The death flow
3. The flow for new (incident) infections
4. The aging flow representing the transition from one age group to its immediately older group
5. The first-order delay flow based on the schema of the system structure diagrams

To support a clearer visualization, these flows are drawn using five different colors. The type system S_{type} has five auxiliary variables corresponding to these five flows—but for simplicity, we do not depict these auxiliary variables. It also has three sum auxiliary variables:

1. N , representing the total population of the whole model
2. NS , representing the population of a specific subgroup
3. NI , representing the count of infectious persons of a specific subgroup

Finally, the type system S_{type} has nine links.

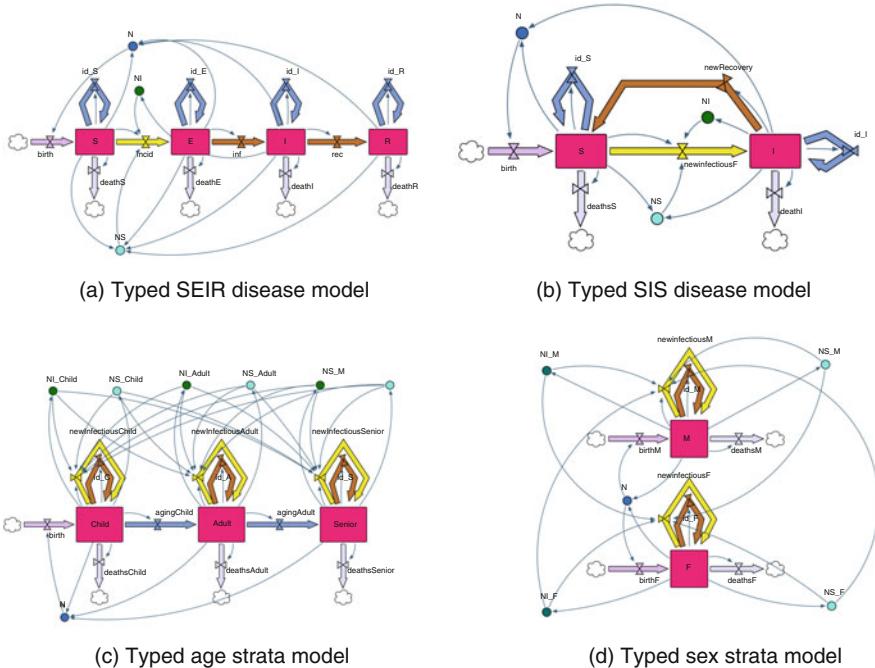


Fig. 8.13 Four examples of typed stock-flow diagrams

Figure 8.13 shows four system structure diagrams “typed” by S_{type} : that is, equipped with maps to S_{type} . The first two are infectious disease models which can serve as aggregate models $S_{\text{aggregate}}$: an SEIR model and an SIS (Susceptible–Infectious–Susceptible) model. The second two can serve as strata models S_{strata} : a sex strata model and an age strata model. The “typing” of these four stock-flow diagrams—that is, their maps to S_{type} —is indicated by colors following the coloring scheme of Fig. 8.12.

In Fig. 8.14 we show the results of pullback-based stratification for each combination of an aggregate model (either the SEIR model or SIS model) and a strata model (either the age strata model and sex strata model). A stratified model $S_{\text{stratified}}$ is generated by taking the pullback of each of the four combinations of an aggregate model and a strata model. These four stratified models are the age-stratified SEIR model, age-stratified SIS model, sex-stratified SEIR model, and sex-stratified SIS model.

More generally, we can define many other system structure diagrams $S_{\text{aggregate}}$ to serve as aggregate models for infectious diseases [2]. Similarly, we can define many different system structure diagrams S_{strata} to serve as strata models characterizing the structure and patterns of progression associated with different types of stratification: not only by sex and age but also by socioeconomic or employment status, and geographical stratification including mobility amongst regions, mobility amongst

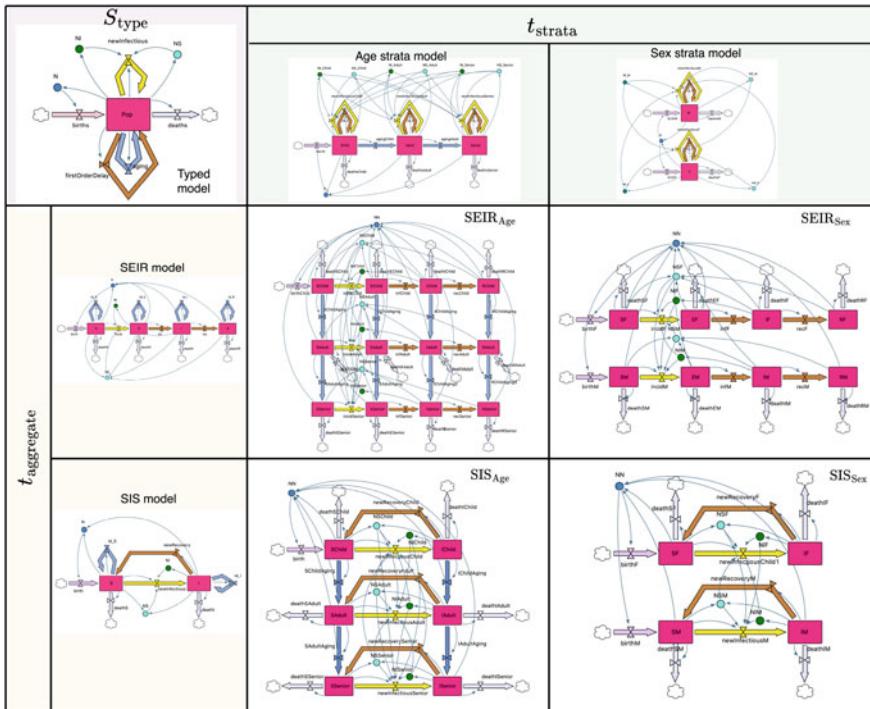


Fig. 8.14 Examples of four stratified models

regions from a home base in a particular region, etc. Once an aggregate model and strata model are chosen along with their typings $t_{\text{aggregate}}: S_{\text{aggregate}} \rightarrow S_{\text{type}}$ and $t_{\text{strata}}: S_{\text{strata}} \rightarrow S_{\text{type}}$, our code can automatically construct a stratified model by taking a pullback. For example, we can generate an SEIR age-stratified system structure diagram by calculating the pullback of the SEIR model and an age strata model.

Moreover, we can build stratified models with multiple dimensions by taking repeated pullbacks of multiple stock-flow diagrams. For example, we can build an age-and-sex-stratified SEIR model by such an iterated pullback involving strata models for each of two dimensions—sex and age. The result is shown in Fig. 8.15. Similar approaches can be used to model progression of multiple comorbidities and behavioral risk factors—a form of stratification that, done by hand, would be subject to a combinatorial explosion of detail [23].

As mentioned, the stratified models here are built based on system structure diagrams, with the goal of simplifying the stratification process and avoiding the need to consider the functions in the stock-flow diagrams. Fully stratified stock-flow diagrams are then generated by assigning functions to each auxiliary variable of the stratified system structure diagram. Figure 8.16 shows a solution of an example SIS model stratified by sex.

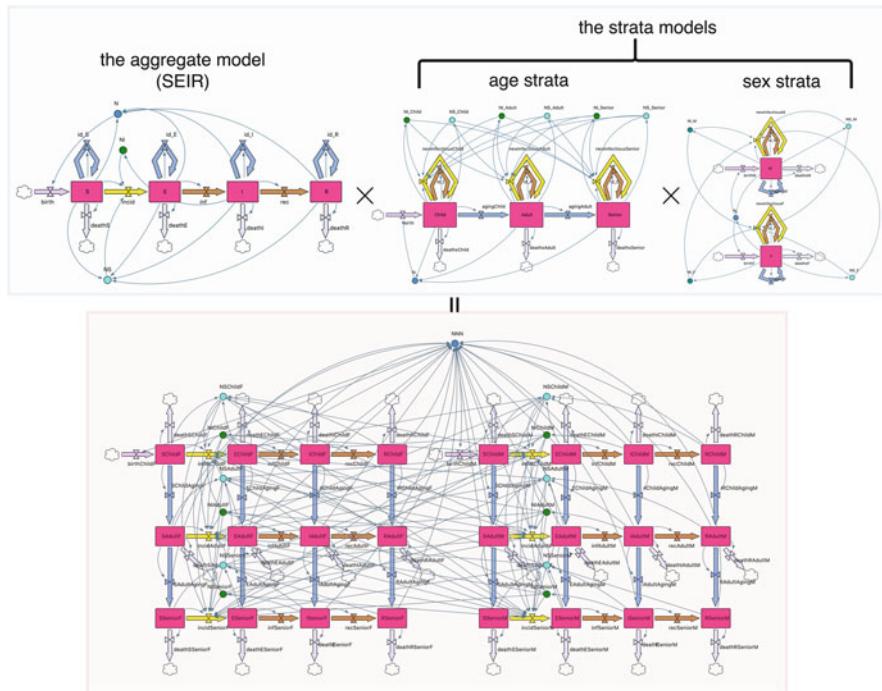


Fig. 8.15 Example of building an SEIR model stratified in multiple dimensions

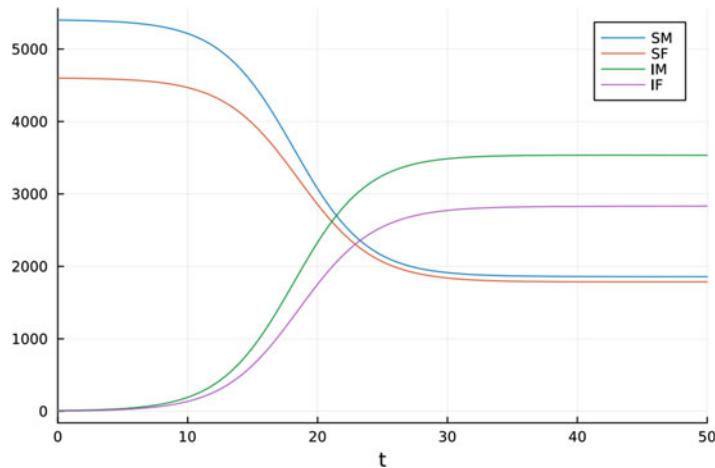


Fig. 8.16 Solutions of the ODE semantics for a sex-stratified SIS model

8.6 ModelCollab: A Graphical Real-Time Collaborative Compositional Modeling Tool

While the modular modeling approaches explored in this chapter make ubiquitous use of category theory, use of the resulting functionality—for example, the ability to compose and stratify diagrams or to interpret them through differing analyses—does not require knowledge of that foundation. Impactful infectious disease modeling is typically conducted in interdisciplinary teams, and securing timely, ongoing feedback about model structure and emergent behavior from non-modeler team members is key to both model refinement and organizational learning from modeling. Often tacit knowledge of non-modeling team members concerning the system under study (say, evidence for episodic reemergence of a communicable disease in certain demographic segments despite prevention and control efforts) is only elicited once team members have a chance to comment on visualizations of model structure and summaries of model dynamics. Often interpretation of such dynamics is greatly enhanced through reasoning about the relationship between observed behavior and the diagram structure—for example, through recognizing that an increase in a variable over time reflects a situation where inflow is greater than outflow, explaining an invariant value of a state variable in terms of a balance between inflows and outflows, or reasoning about exponential change in terms of driving feedback loops.

Partly for these reasons, the system dynamics tradition of modeling has long prized the use of visual modeling software that keeps the attention of modelers—and other team members—on diagrams depicting model structure. While such software does support communication across interdisciplinary team members, it suffers both from the disadvantages of traditional treatment of such models discussed in the introduction and a limit to being modified—and often viewed—by only a single user at a time.

Here we describe open-source, visual, collaborative, categorically rooted, diagram-centric software for building, manipulating, composing, and analyzing system dynamics models. This web-based software, named ModelCollab, is designed for real-time collaborative use across interdisciplinary teams. The graphical user interface allows the user to interactively conduct the types of categorically rooted operations discussed without any knowledge of their categorical foundations. This software is at an early stage and currently supports only a subset of the options made possible by the StockFlow.jl framework on which it rests. But its development is progressing rapidly, and we anticipate an expansion to eventually handle a far larger set of operations. We describe the use of some of the early features of this system, bearing in mind that multiple users will commonly be using the system simultaneously.

ModelCollab provides a modal interface for adding diagram components to a Canvas used to display the diagram being assembled. Different concurrent users can be present in different modes at the same time. For example, within “Stock” mode, the user can click on the canvas to add a Stock and similarly for “Flow,”

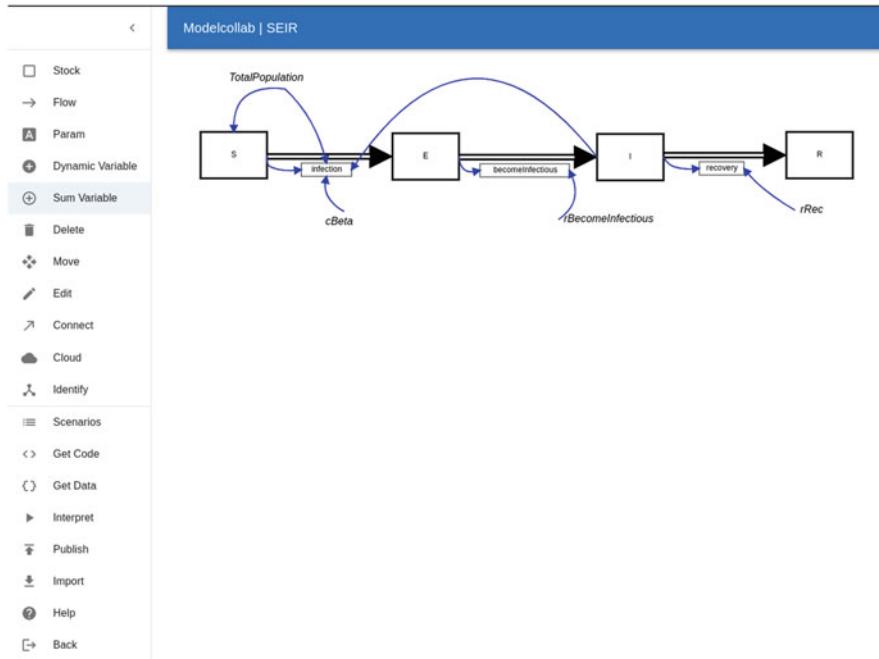


Fig. 8.17 An SEIR diagram built in ModelCollab

“Auxiliary Variable,” “Sum Variable” modes. “Connect” mode is used to establish links indicating dependencies, such as those for auxiliary and sum variables. The interface abstraction level sometimes exceeds that of StockFlow.jl; for example, flows within the graphical interface are shown as depending directly on other variables in the diagram, rather than only via connections with a distinct auxiliary variable. Through this interface, larger diagrams can be created; for example, Fig. 8.17 depicts an SEIR diagram built in the system.

As in online collaborative software such as Google Docs, diagrams can be accessed on an ongoing basis by multiple parties once they have been created in ModelCollab. It is also possible to persist diagrams in other forms within the system. The interface allows export of diagrams through a menu item, with that diagram then being downloaded to the invoking user’s local computer as a JSON [8] file. Beyond exporting, the system offers a structured means of “publishing” diagrams to a simple “Diagram Library” once they have attained a sufficient level of maturity to be worth sharing. Such published diagrams can then be reused by others.

A foundationally important component of ModelCollab functionality is the ability to compose diagrams. Like other diagram assembly operations within ModelCollab, this operation is performed graphically. As a first step, the user can add previously published diagrams into the model canvas, where this imported

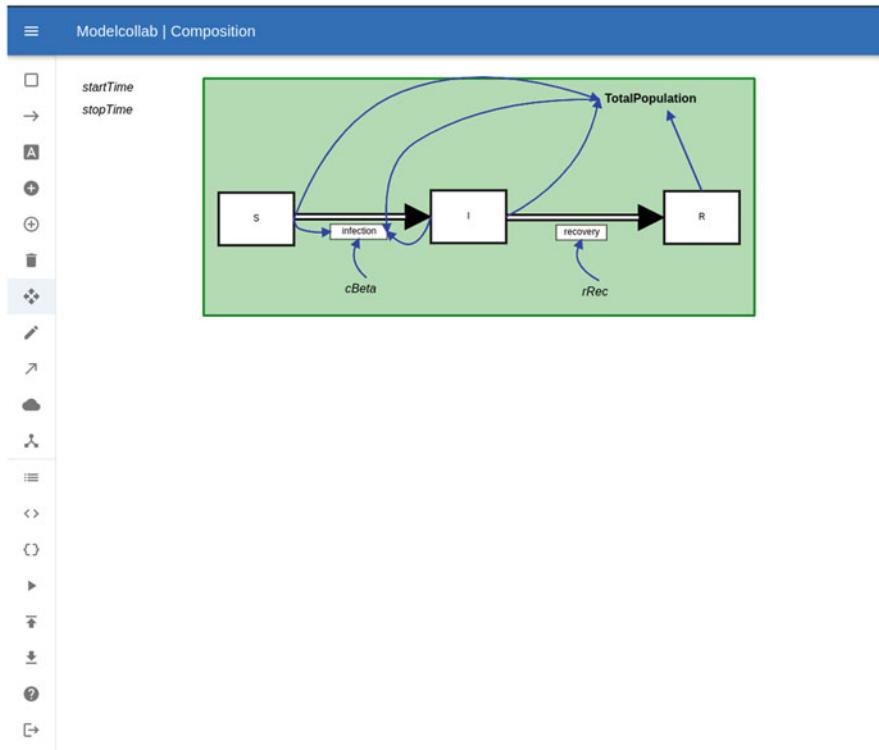


Fig. 8.18 ModelCollab: results of importing a single previously published diagram

diagram (henceforth referred to as the “subdiagram”) is visually distinguished from the surrounding diagram being assembled, as is depicted in Fig. 8.18.

More than one such published diagram can be imported, in which case each is distinguished by a different color. For example, Fig. 8.19 shows the results of importing a second subdiagram, which lies beneath the first imported subdiagram. Beyond these subdiagrams, the canvas will commonly contain a surrounding diagram.

Just after being imported, the two subdiagrams are independent of one another; while this is sometimes appropriate, often the user will recognize ways in which the process depicted in a specific imported subdiagram is coupled with the processes depicted by the surrounding diagram or in the other imported subdiagrams. Such coupling is represented by *composition* of the diagrams via the interface. Through user interface actions, a user can elect to identify (unify) a stock or sum dynamic variable with a variable of the same type in a subdiagram. Such identification of pairs of variables may be performed between variables in an outer canvas diagram and in a subdiagram, or (alternatively) between two variables in different subdiagrams. For example, Fig. 8.20 graphically illustrates the results of identification of

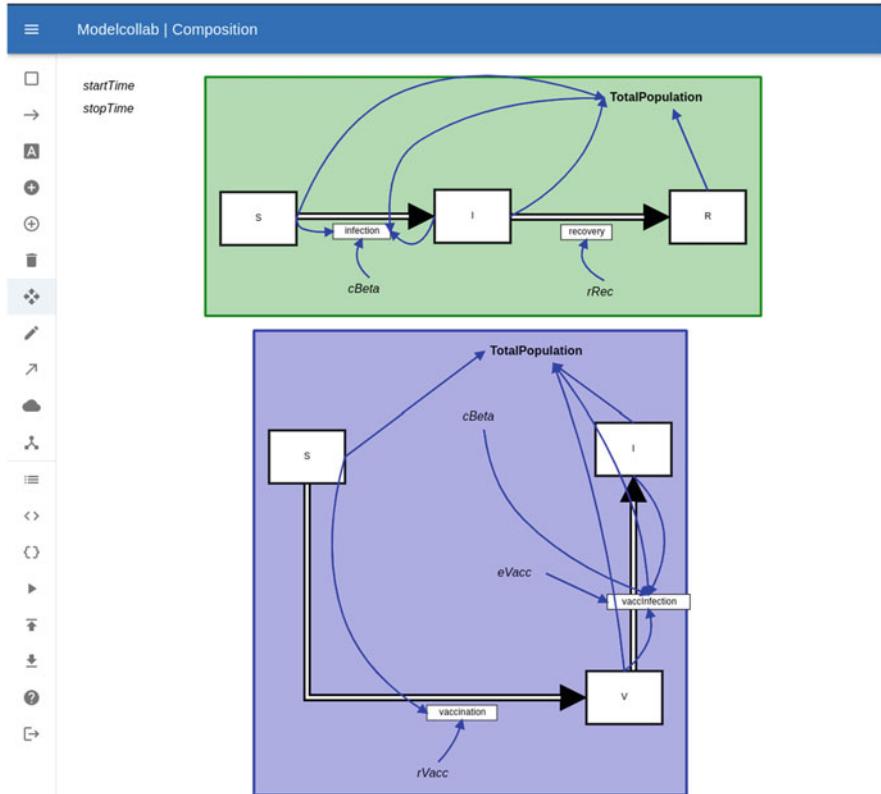


Fig. 8.19 ModelCollab: results of importing a second diagram

a stock V of the upper and lower diagrams. By entering “Identify” mode from the ModelCollab menu, the user can indicate with a pair of successive clicks the pair of stocks to be identified.

Figure 8.21 shows the results of using the “Identify” mode of ModelCollab to identify not just additional stock I between the two diagrams, but also sum variable N .

Once built, ModelCollab supports rendering the definition of a diagram in code form. Specifically, the system offers a “Get code” menu item that produces a file containing Julia code to create the current diagram using calls to StockFlow.jl. If desired, such code could then be used to interactively manipulate the models from a Julia codebase or within a Jupyter Notebook.

Beyond operating on syntactic constructs in the form of diagrams, ModelCollab provides an interface to interpret diagrams according to a menu-chosen semantics. At the time of writing, the software only supports ODE semantics (see Fig. 8.22), but implementation of the other semantics discussed in Sect. 8.3 is planned within the near future.

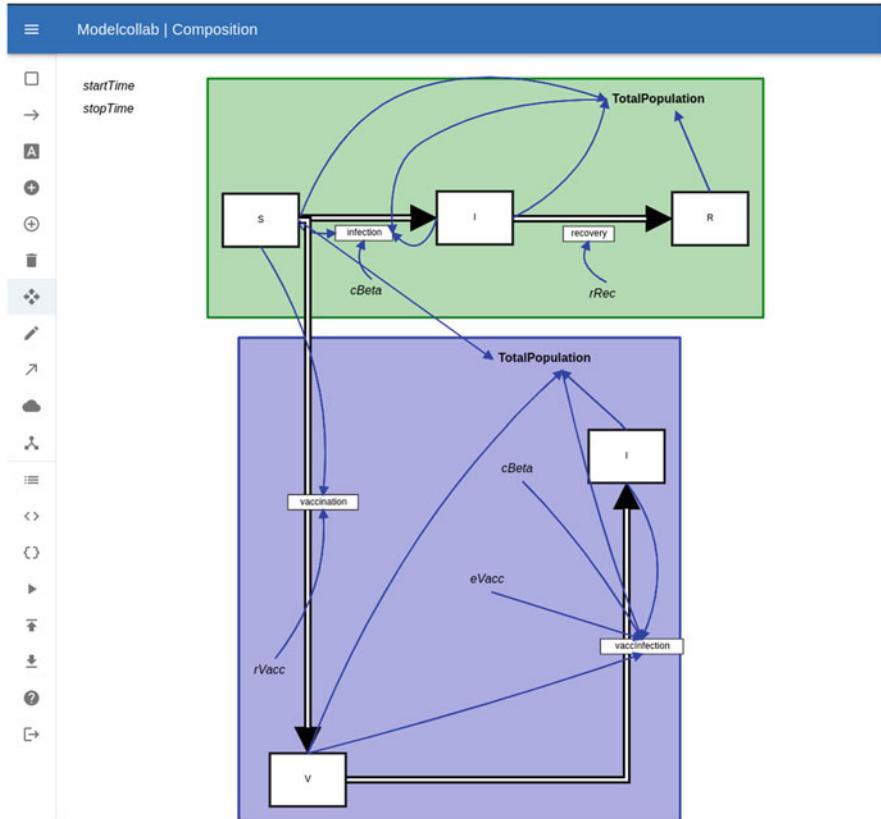


Fig. 8.20 ModelCollab: example of two subdiagrams composed by identifying stock S

Interpretation of a diagram by ODE semantics involves numerical integration of that diagram over a user-specified timeframe. Outputs from that simulation are currently rendered and downloaded as a PNG image file; Fig. 8.23 shows an example.

The current ModelCollab software represents a modest step towards realizing the promise of compositional approaches for the broader modeling community. There are several priorities anticipated for rollout in the near future. Putting aside additions requiring changes for StockFlow.jl itself (some of which are discussed below), planned features include support for pullback-based stratification and additional semantic domains. Furthermore, a scenario-based interface is being planned that will provide persistent options to interpret the model using different semantics and settings, and support other users in observing and annotating the output from semantic-based interpretation long after it is first produced. Support is also planned for the type of seamless version control standard in many real-time collaborative systems and indications of the presence of the pointers of different concurrent users.

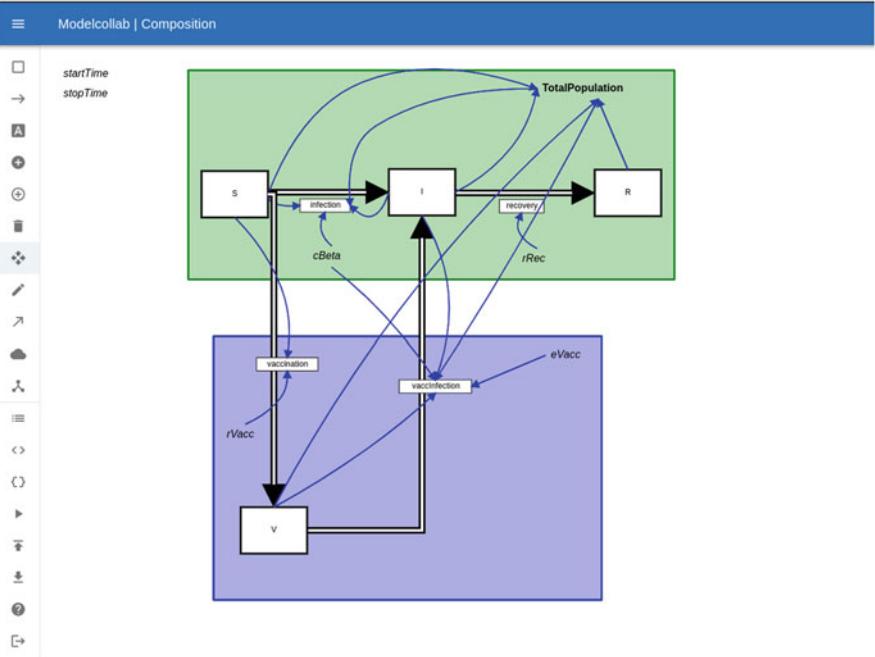


Fig. 8.21 ModelCollab: example of two subdiagrams composed by identifying both stocks S and I and sum variable N

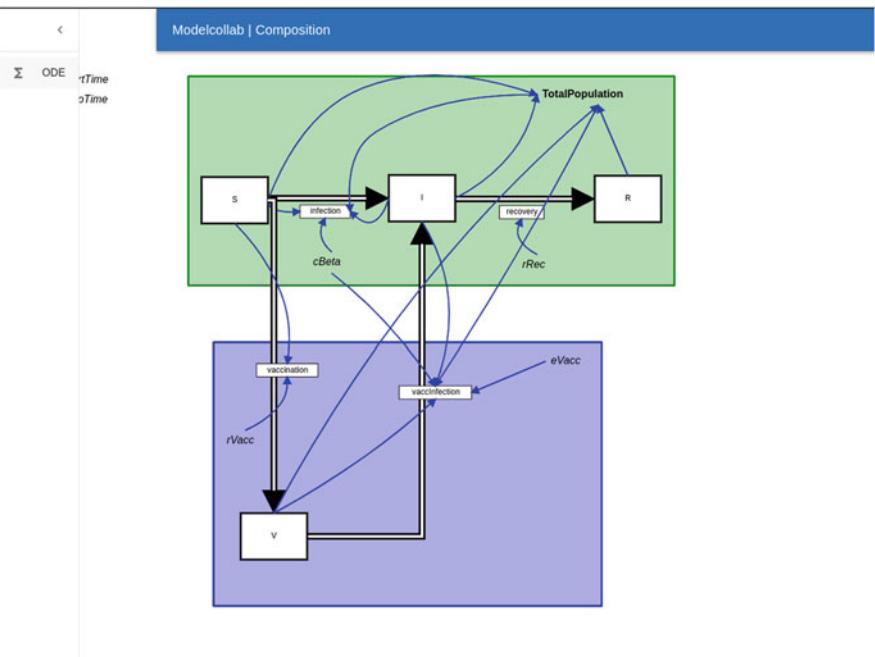


Fig. 8.22 ModelCollab: example of using the “Semantics” menu to elect to interpret the diagram (the syntax) as an ODE

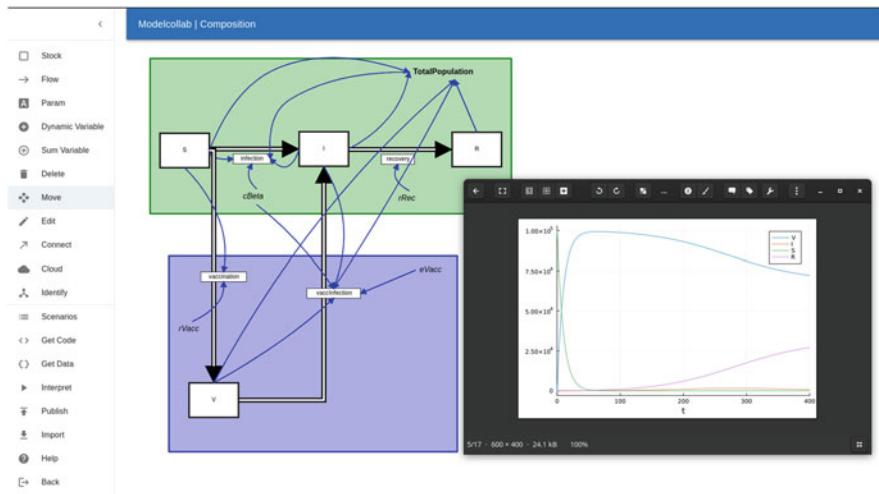


Fig. 8.23 ModelCollab: example of results of interpreting the composed diagrams (the syntax) with ODE semantics

Finally, role-based security, authentication, and authorization systems are planned to formalize permission-based enablement of diagram use, sharing, and modes of interaction.

8.7 Conclusion

This chapter demonstrates some of the modeling benefits secured when one takes diagrams seriously. Specifically, we have offered a brief look at some of the characteristics of categorical treatment of stock-flow diagrams, while offering a nod to other related diagrams within the diagram-centric system dynamics modeling tradition. While our treatment has been informal and has only touched on a small subset of the consequences of a categorical foundation for diagrams, we have highlighted several benefits: the capacity to promote modularity and reuse via diagram composition, to sidestep model opacity arising from the curse of dimensionality afflicting stratified models via a modular stratification, and better supporting the needs of interdisciplinary stakeholders via the capacity to interpret the same diagrams through the lens of varying semantic domains. Some of these benefits—such as those involving semantics mapping stock-flow diagrams to system structure diagrams or causal loop diagrams—describe relationships between different types of diagrams not previously formalized. Other outcomes of providing a firm mathematical basis for stock-flow models have been noted in passing. For example, such a formalization offers the ability to soundly transform a diagram whilst maintaining invariant its mathematical meaning, such as for optimization

or parallelization of model simulation. As another example, the formalization can also allow the use of maps between diagrams that coarse-grain model structure. There are diverse other opportunities for exploiting this categorical formalization of mathematical structure.

While it offers promise, this work remains at the earliest stages of exploiting such opportunities extending from categorical stock-flow diagrams. StockFlow.jl and ModelCollab require many important extensions to substantively address the challenges of practical modeling. Key priorities include providing support for upstream/downstream composition based on “half-edge” flows emerging from a diagram, supporting methods for hierarchical composition of diagrams (such as those pioneered by our colleague N. Meadows [22]), adding full support for causal loop diagrams and system structure diagrams, and supporting the augmentation of multiple types of diagrams with dimensional information [14] and use of such information in stock-flow diagram composition, stratification, and additional semantic domains. There is further a need and opportunity to develop additional semantic domains for use with stock-flow models. These include those associated with different numerical simulations, such as stochastic transition systems, stochastic differential equations, and difference equations, as well as those associated with computational statistics techniques. We seek to follow each such advance in the formalisms for stock-flow models by successive implementation in StockFlow.jl and ModelCollab.

At a time where health dynamic modeling is in greater demand and more needed than ever, formalizing its categorical foundation confers benefits key to addressing the team-based modeling needs of the twenty-first century. We enthusiastically welcome collaboration with colleagues interested in exploring opportunities to transformationally enable health modeling by tapping the power of category theory.

Acknowledgments We gratefully acknowledge the extensive insights, comments, and feedback from Evan Patterson of the Topos Institute. Co-author Osgood wishes to express his appreciation of support via NSERC via the Discovery Grants program (RGPIN 2017-04647), from the Mathematics for Public Health Network, and from SYK & XZO.

References

1. AlgebraicJulia: Bringing compositionality to technical computing. <https://www.algebraicjulia.org>
2. Anderson, R.M., May, R.M.: *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford (1992)
3. Andrieu, C., Doucet, A., Holenstein, R.: Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. B (Statistical Methodology)* **72**(3), 269–342 (2010)
4. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **50**(2), 174–188 (2002)
5. Baez, J.C., Courser, K.: Structured cospans. *Theor. Appl. Categ.* **35**(48), 1771–1822 (2020). [arXiv:1911.04630](https://arxiv.org/abs/1911.04630)

6. Baez, J.C., Courser, K., Vasilakopoulou, C.: Structured versus decorated cospans. *Compositionality* **4**(3) (2022). [arXiv:2101.09363](https://arxiv.org/abs/2101.09363)
7. Baez, J.C., Li, X., Libkind, S., Osgood, N.D., Patterson, E.: Compositional modeling with stock and flow diagrams. To appear in Applied Category Theory 2022, Electronic Proceedings of Theoretical Computer Science (2022). [arXiv:2205.08373](https://arxiv.org/abs/2205.08373)
8. ECMA-404: The JSON data interchange standard. <https://www.json.org/json-en.html>
9. Fong, B.: Decorated cospans. *Theor. Appl. Categ.* **30**(33), 1096–1120 (2015). [arXiv:1502.00872](https://arxiv.org/abs/1502.00872)
10. Fong, B.: The Algebra of Open and Interconnected Systems. Ph.D. Thesis, Computer Science Department, University of Oxford (2016). [arXiv:1609.05382](https://arxiv.org/abs/1609.05382)
11. Fong, B., Spivak, D.I.: Hypergraph categories (2018). [arXiv:1305.0297](https://arxiv.org/abs/1305.0297)
12. Fong, B., Spivak, D.I.: An Invitation to Applied Category Theory: Seven Sketches in Compositionality. Cambridge University Press, Cambridge (2019)
13. Gelb, A.: Applied Optimal Estimation. MIT Press, Cambridge (1974)
14. Hart, G.W.: Multidimensional Analysis: Algebras and Systems for Science and Engineering. Springer, Berlin (1995)
15. Hovmand, P.S.: Community Based System Dynamics. Springer, Berlin (2014)
16. Lawvere, F.W.: Functorial semantics of algebraic theories. *Proc. Natl. Acad. Sci.* **50**(5), 869–872 (1963)
17. Leinster, T.: Basic Category Theory. Cambridge University Press, Cambridge (2014). [arXiv:1612.09375](https://arxiv.org/abs/1612.09375)
18. Li, X., Doroshenko, A., Osgood, N.D.: Applying particle filtering in both aggregated and age-structured population compartmental models of pre-vaccination measles. *PLoS ONE* **13**, e0206529 (2018)
19. Li, X., Keeler, B., Zahan, R., Duan, L., Safarishahrbijari, A., Goertzen, J., Tian, Y., Liu, J., Osgood, N.D.: Illuminating the hidden elements and future evolution of opioid abuse using dynamic modeling, big data and particle Markov chain Monte Carlo (2018)
20. Libkind, S., Baas, A., Halter, M., Patterson, E., Fairbanks, J.: An algebraic framework for structured epidemic modeling. *Philos. Trans. R. Soc. A.* **380**(2233), 20210309 (2022). [arXiv:2203.16345](https://arxiv.org/abs/2203.16345)
21. Libkind, S., Baas, A., Patterson, E., Fairbanks, J.: Operadic modeling of dynamical systems: mathematics and computation. *EPTCS* **372**, 192–206 (2022). [arXiv:2105.12282](https://arxiv.org/abs/2105.12282)
22. Meadows, N.: Hierarchical composition of stock & flow models. CEPHIL Technical Report (2022)
23. Osgood, N.: Representing progression and interactions of comorbidities in aggregate and individual-based systems models. In: Proceedings of the 27th International Conference of the System Dynamics Society. Albuquerque, New Mexico (2009)
24. Osgood, N.D., Eng, J.: Effective use of pmcmc for daily epidemiological monitoring and reporting: methodological lessons. Abstract & Conf. Publication, Ann. Meet. of the Statistical Society of Canada (2022)
25. Osgood, N.D., Liu, J.: Towards closed loop modeling: Evaluating the prospects for creating recurrently regrounded aggregate simulation models using particle filtering. In: Proceedings of the 2014 Winter Simulation Conference, WSC '14, pp. 829–841. IEEE Press, Piscataway (2014)
26. Patterson, E., Lynch, O., Fairbanks, J.: Categorical data structures for technical computing. *Compositionality* **4**(2) (2022). <https://doi.org/10.32408/compositionality-4-5>
27. Qian, W., Osgood, N.D., Stanley, K.G.: Integrating epidemiological modeling and surveillance data feeds: a Kalman filter based approach. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, pp. 145–152. Springer, Berlin (2014). https://doi.org/10.1007/978-3-319-05579-4_18
28. Safarishahrbijari, A., Teyhouee, A., Waldner, C., Liu, J., Osgood, N.D.: Predictive accuracy of particle filtering in dynamic models supporting outbreak projections. *BMC Infect. Dis.* **17**(1), 1–12 (2017)

29. Spivak, D.I.: The operad of wiring diagrams: formalizing a graphical language for databases, recursion, and plug-and-play circuits (2013). [arXiv:1305.0297](https://arxiv.org/abs/1305.0297)
30. Spivak, D.I.: Category Theory for the Sciences. MIT Press, Cambridge (2014). Preliminary version. [arXiv:1302.6946](https://arxiv.org/abs/1302.6946)
31. Sterman, J.D.: Business Dynamics. McGraw-Hill, New York (2000)
32. Stockflow.jl. <https://github.com/AlgebraicJulia/StockFlow.jl>

Chapter 9

Agent-Based Modeling and Its Trade-Offs: An Introduction and Examples



G. Wade McDonald and Nathaniel D. Osgood

9.1 Introduction

Agent-based modeling (ABM) is a computational dynamic modeling technique which seeks to understand the behavior of complex systems through the lens of agent-agent and agent-environment interactions [44]. Agent-based models (ABMs) can be said to be “upward-facing” or “bottom-up” in the sense that we specify behavior of situated agents and these agents interact, which dictates higher-level system behavior. Patterns and often surprising results emerge over time, space, and networks, possibly at multiple levels of the system.

9.2 Characteristics of Agent-Based Models

The origins of agent-based modeling can be traced back to von Neumann and Ulam’s work in the 1940s on replicating and cellular automata [31, 56]. Since that time, the tradition has been enriched by contributions from computational physics, computer science, and mathematics and from microsimulation modeling in economics. ABMs, like other types of dynamic models, vary from the extremely stylized and simple thinking tools for theory building [18] to descriptively rich and empirically grounded models that seek to support theory explication and understanding the logical consequences of theory over time [18].

G. W. McDonald (✉) · N. D. Osgood
University of Saskatchewan, Saskatoon, SK, Canada
e-mail: gwm762@mail.usask.ca; osgood@cs.usask.ca

ABMs consist of one or more populations of agents, where each such agent is equipped with parameters (representing prespecified assumptions), state (characterizing an underlying situation evolving over time), and actions that change that state according to some rules or rates of change. These models are specified over some time horizon according to either a continuous or discrete time abstraction. *Continuous time modeling* abstractions support discrete events occurring at real-valued times at whatever tempo, pace, or temporal granularity is required for particular circumstances within the model. For example, such an event might be associated with each occurrence of infection, recovery, vaccination, contact, death, or birth. By contrast, the *discrete time abstraction* involves updates to model agents and environments in lockstep in monolithic (atomic) “ticks” or “timesteps”; in the event that there are multiple processes that need to be considered within the timespan represented by a given such timestep, their effects need to be brokered in the associated update to model step. For example, each timestep might represent a month as a whole, and on reaching that timestep, all of the distinct processes occurring during that month (deaths, infections, births, vaccinations) would need to be considered.

Beyond having properties of its own, each agent is situated in an environment, which can include static or dynamic networks, spatial context that may be geographic or stylized, and potentially several levels of context. Sometimes such environments are highly evidenced and empirically grounded. Beyond these, there are typically some outputs of model state or changes therein reported by the model; sometimes those are governing factors in the model, which are often instead epiphenomenal—that is, reporting on but not influencing model state evolution. Finally, interventions represent mechanisms that alter elements of a model to permit investigation of counterfactual scenarios.

9.2.1 Parameters

Figure 9.1 shows an example of a population of agents within an agent-based model, with each agent representing a person and having parameters for sex, income, and ethnicity. Values of these parameters are fixed over time but vary from agent to agent within the population. Agent-based models commonly include parameters that are continuous (e.g., income) and discrete (e.g., sex) in character; they can also be relational—such as a parameter referring to an agent’s mother or school. In this context, *continuous* refers to a parameter that may hold any value on an interval on the real number line (as approximated by floating-point arithmetic), while *discrete* refers to a parameter that may take on a limited number of distinct values, whether numeric, ordinal, or nominal. This capacity of agent-based models to capture continuous and relational heterogeneity stands in contrast to the fact that the only general approach to representing heterogeneity in aggregate models—via model stratification—is limited to representing discrete heterogeneity. A further advantage of individual-level representation—the capacity to scale far more effectively than

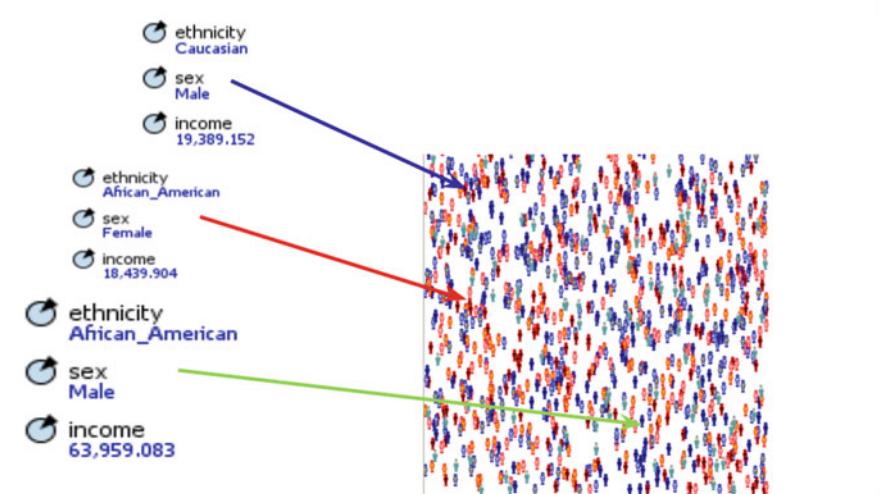


Fig. 9.1 An example of agent parameters: Parameter values are constant or otherwise prespecified, but they do vary from agent to agent across a population

aggregate models as the number of types of heterogeneity rises and to nimbly evolve the types of heterogeneity represented—is also shared by aspects of state representation and will be discussed below.

9.2.2 State, Actions, and Rules

A number of ABM software packages—including the AnyLogic software [54] used for the example models here—describe agent state, and the actions and rules by which it evolves, using statecharts. Figure 9.2 shows an example statechart for an agent representing a person in an infectious disease model. A person is associated with a set of possible states related to infection status indicating that at any one time they are either susceptible, exposed, infective, or recovered. Over time, they evolve between those states. The statechart at once depicts the possible states as rounded rectangles as well as the actions that can change state as arrows—for example, transitioning from latent infection to infectiousness. The transition internal to the infective state is associated with this agent’s exposure of other agents to pathogen. The iconography on the arrows hints to the fact that there are rules of various types governing these actions. Within one statechart, states are mutually exclusive and collectively exhaustive. Unless explicitly coupled, multiple statecharts within one agent evolve independently.

When compared to aggregate dynamic modeling methods, the representation of state in agent-based models confers some advantages. We can readily represent a given agent’s state and its evolution over time with respect to more than one

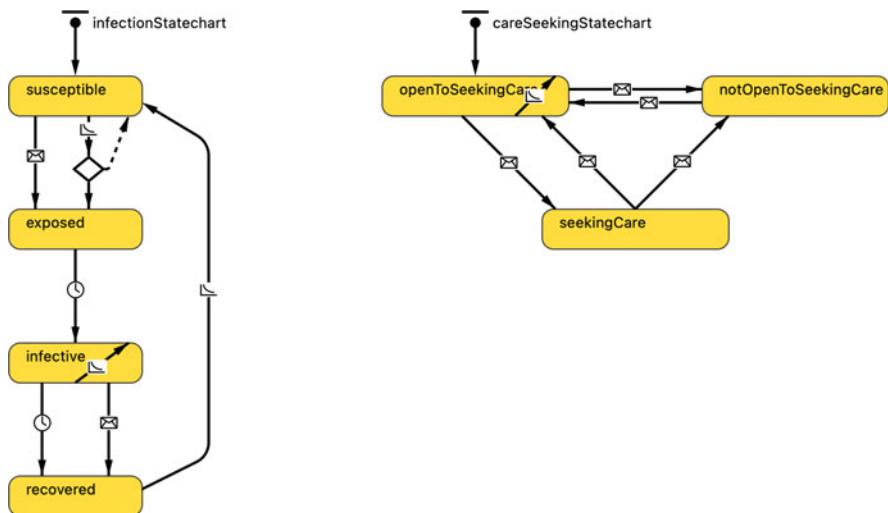


Fig. 9.2 Examples of statecharts in a single agent: Rounded rectangles represent states while arrows represent transitions, which encompass the actions and rules governing state change over time

type of concern—for example, we can represent a person as in a particular state with respect to infection and a particular state with respect to care-seeking. Such a representation avoids the “curse of dimensionality” that confronts stratified aggregate models as a result of the combinatorial explosion of possible states of different conditions. The count of such combinations rises geometrically with the number of dimensions being represented [34, 36]. For the example in Fig. 9.2, there are two statecharts with a total of seven states; representation of the same level of heterogeneity using aggregate compartmental modeling techniques would require $4 \times 3 = 12$ compartments; similar scaling is observed for static aspects of heterogeneity. The capacity to separate such concerns into separate statecharts (or, for static heterogeneity, into different parameters) further allows changes in the number or design of such statecharts (or in the types of heterogeneity) to be performed in a modular, localized fashion; this contrasts with the global changes that are required across the scope of an aggregate model as dimensions of heterogeneity or dimensions of progression are changed.

Finally, within an agent-based model, there is no requirement that statecharts be fully independent of one another; we can allow them interact in ways that are defined through localized interactions, but otherwise evolve fairly independently. Such interactions can be implemented in an elegant fashion compared to the combinations that are required for highly stratified aggregate models.

9.2.3 Environment

Agents are not solitudes; they are placed in an environment which situates them in some context. This context commonly situates agents in space and/or in networks with connections to other agents. A person can have their presence represented in one or more networks, for example, a family network, collegial network, social network, intravenous drug use network, or sexual network. While network connections between agents of the same type are common (e.g., between two persons), so are connections between agents of different types—for example, networks connecting persons and community service providers, or population members and their physicians. Often connections within networks serve as conduits for interactions between the pair of connected agents. The most common mode of interaction over such connections—and, by extension—and networks is via message passing where one agent sends a message to another. Such message passing provides a very flexible and computationally elegant means of characterizing agent interactions along one or more networks.

Beyond networks, we also often place agents within a spatial context. Diverse types of spatial environments can be found in the literature, including 2D or 3D Euclidean, irregular, toroidal, discrete square, triangular, or hexagonal lattices and geographic spaces. Figure 9.3 shows an example of a model featuring an irregular spatial environment, Fig. 9.4 shows an example with a 3D environment, and Fig. 9.5 shows an example with a stylized discrete square lattice spatial environment.

There are diverse motivations for placing agents in spatial environments. Certain types of spatial environment are recommended for certain needs. With a geographic environment, we can capture effects involving locality, perception, and influence in an empirically situated spatial context. This can capture, for example, aspects of social determinants of health, like the presence of food deserts, areas with



Fig. 9.3 Example of an agent-based model employing an irregular spatial environment [54]



Fig. 9.4 Example of an agent-based model employing a 3D spatial environment



Fig. 9.5 Example of an agent-based model employing a stylized discrete square lattice spatial environment

high nitrous oxide levels, or areas underserved by health-care provision. Perhaps we're interested in behavior that is actually spatial in nature, like mobility changes resulting from investments in walkability or support groups for walking. Or, perhaps we're interested in disparities in COVID-19 hospitalization burden that result from the clustering of unvaccinated people or of individuals at high risk of infection due

to high chronic disease burden. Situating agents in geographic context can allow us to capture the disproportionate risks in certain regions. Using models featuring spatial context allows us to look beyond an average burden of a disease or policy and consider key local variations—often pockets where outbreaks or policy resistance occurs where average measures would suggest that none should occur.

9.2.4 Outputs and Emergent Behavior

A key need with all models is to understand their behavior over time. We often are keenly interested in emergent behavior evinced in model outputs; sometimes those emergent behavior patterns can be quite eye-opening or surprising. Sometimes these outputs correspond to factors that govern the evolution of model state. In other cases, we examine epiphenomenal factors that characterize some aspect of model state (often serving as summaries of sorts), but do not drive it. Like aggregate models, commonly the behavior of an agent-based model is considered as a function of time. However, agent-based models offer a wider repertoire of behaviors, for ABMs may output measures of state or behavior over networks or space—for example, reporting where infection or risk concentrate in particular regions of the network or in a geographic region. Sometimes, seeing such patterns can offer great value in understanding model behavior. For example, within a chronic wasting disease model [29], the concentration of risk of exposure to prions in areas near the water margin may shape infection risk in ways that manifest in unexpected impacts on the population over time.

9.2.5 Stochastics

While aggregate models routinely abstract away from particular events, and focus instead on broad patterns [46, 47], most ABMs deal with individual-level events, such as exposure to infection, recovery, events associated with care-seeking, or decisions as to where to seek out food. When depicting factors at the level of individual events, it is common that health ABMs consider elements of human behavior and psychology that—given suitable model scope—are best described as stochastic. Treating such behaviors as subject to stochastic evolution does confer advantages, such as the ability to allow us to explain variability we see in real-world data, but also places an onus on us as modelers to ensure that observed model results are not merely flukes resulting from one chance event or series of events in a model but instead reflect regularities and structure within the behavior of the system being modeled. Critical to offering well-founded scientific insights, we must ensure that results are replicable. In order to achieve this, we typically run a model many times over and examine results from those many realizations, called an *ensemble*.

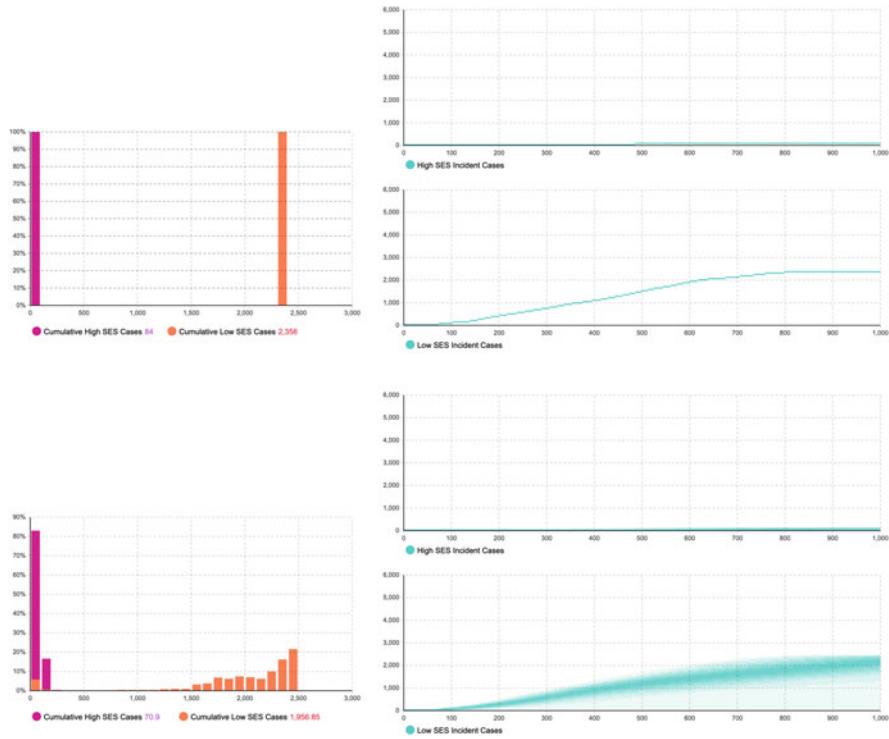


Fig. 9.6 Example of output from a single run of an agent-based model compared with that from an ensemble

If we consider the example in Fig. 9.6, we can see the difference between a single realization of a stochastic model and an ensemble. A single run of this model reveals a marked difference in cases between the two socioeconomic status (SES) groups, with a higher case count for the low-SES group. If we consider an ensemble of many runs, it is revealed that the initial single run represents just one of many possible outcomes. The ensemble reveals that randomness in the initial state together with stochastics in the model induces empirical distributions on the case counts for each of the low- and high-SES groups. Over time, we see some realizations with smaller numbers of cases, some with larger, and some where population-wide spread never takes off in either the high- or low-SES group.

9.2.6 *Interventions*

Many—but not all—of the motivation to use ABMs focus on “what-if” questions, reflecting a desire to evaluate interventions and counterfactuals. Within such scenarios, we seek to ask the model: “Given the assumptions of this model, what would

be the impacts of undertaking action X ?" One of the most powerful things about ABMs is that they can be used to examine the effects of intervention mechanisms exploiting *individual-level longitudinal information*—with the way in which the intervention acts upon a given person depending on, for example, a person's history, in addition to their static characteristics and current state. An intervention examined using an ABM could be targeted based on, for example, a person's past presentation history or episodes of care—for example, with a sexually transmitted infection (STI) clinic or dental office offering behavioral counseling for individuals whose recent presentation history suggests risk behavior, or individuals with recent acute hospitalizations could be targeted for regular dual COVID-19/influenza vaccination. Beyond depending on individual history, intervention design and implementation can also exploit ABM's rich capacity to represent heterogeneity and context by basing intervention mechanisms on individual characteristics or based on geographic or network position. As an example, such models could represent and evaluate impacts of provisioning distinct care pathways to ensure culturally appropriate care for members of indigenous population. As another example, control and prevention of STIs recognized the priority of focusing on core groups within STI networks, which is a key priority, and ABMs can readily characterize interventions triggered by or whose details depend on context at any number of different levels within the model which can be very valuable. Or context-targeted interventions could be defined by focusing on addressing neighborhoods situated in food deserts. Other classes of interventions readily characterized in ABMs do not merely target certain contexts, but actively intervene upon them. For example, an intervention may work by *establishing* network connections between people—building social capital or supportive or prosocial influences—as some of the work of Alan Shiell and Penny Hawe [14, 19] and others have investigated. In fact, many interventions focus at a certain level on changing networks—for example, those employing support groups or community-based support organization, accountability partners, and “buddy” systems.

9.3 Example: Chickenpox

The first example model characterizes the dynamics of chickenpox and shingles. This model has been employed to investigate several questions surrounding those diseases [40–42].

9.3.1 *Chickenpox and Shingles*

Chickenpox and shingles are two distinct diseases both caused by the varicella zoster virus (VZV). Chickenpox is typically a childhood disease; once a person has recovered from this disease, the virus generally remains dormant in the body.

VZV can reactivate later in life as shingles, which causes an often debilitating and painful rash suffered by middle- to senior-aged adults [9, 11]. The first question that we investigated when building this ABM was whether vaccination for chickenpox would cause an increase in shingles incidence.

9.3.2 Model Scope

It is often helpful to characterize model scope by describing which features are endogenous to the model, which are exogenous and specified by the modelers, and which are ignored [52]. Endogenous factors in this model include transmission, contact patterns, mother-child dyads, vaccination schedule adherence, fertility, mortality, hospitalizations due to VZV infections, waning of disease-induced immunity, boosting of immunity, and accumulated costs. Implementation of endogenous mechanisms is discussed further in Sect. 9.3.3. Exogenous factors included assumptions about vaccine attitude, the specified vaccine schedule, and assumptions about the effectiveness of the vaccine, which were drawn from clinical research. Additional exogenous factors included initial population demographics as well as population density and unit costs assumed in various cost-effectiveness analyses. Ignored factors include household structure, time variation of contact structures, schools, and childcare facilities. Key parameter values for the chickenpox model are listed in Table 9.1.

9.3.3 Statecharts

The statechart in Fig. 9.7 represents the natural history of disease for the varicella zoster virus. Following a child's birth, they are protected for some months by maternal antibodies; they then become susceptible. Two states represent protection due to vaccination; with only one dose it is possible to have a breakthrough infection while vaccinated, but once two doses are administered, a person is considered to be immune for life. A person may go on to be weakly or fully infected with chickenpox. During the ensuing period of infectiousness, they may transmit the infection to others through exposure messages they send (according to Poisson arrivals) to nearby agents; there is then a period where they continue to show symptoms but are no longer infectious to others. Once a person recovers from chickenpox, they enter the recoveredCP state, within which there is a mechanism that represents a process of episodic boosting of immunity driven by exposure to other people with chickenpox or shingles. Although it has not yet been required in investigating any of our scientific questions that we addressed with the model, the model further characterizes the occurrence of shingles vaccination. The remaining states represent progression of shingles infection, which can be present as a mild case or a severe case, with the latter being designated as postherpetic neuralgia

Table 9.1 Key parameter values for chickenpox model

Parameter	Value(s)	Source
Initial population	500,000	
Mortality and fertility	Multiple	[50, 51]
Initial cell-mediated immunity	<code>max(0.001, normal(0.05, 1))</code>	[33]
Force of reactivation	<code>gamma(2, 0.1, 0)</code>	[33]
Waning of immunity coeff. shingles	0.45–0.93	Calibration
Waning of immunity rate shingles	0.4 year ⁻¹	[33]
Duration of exogenous boosting	0.42–10.0 year	Calibration
Exogenous infection rate	17.83 year ⁻¹	Calibration
Prob. of inf. on contact (normal)	0.78	Calibration
Prob. of inf. on contact (breakthrough)	0.234	[13]
Prob. of inf. on contact (shingles)	0.234	Calibration, [13]
Connection range normal	8.958	Calibration
Connection range preferential	21.245	Calibration, [30]
Preferential contact rate	20	Calibration
Normal contact rate	30.124	Calibration, [30]
Shingles connection range modifier	0.124	Calibration
Preferential mixing age	1–9 years	[24]
Population density urban	0.3	
Population density rural	0.2	
Vaccination attitude	Acceptor = 65%, Hesitant = 30%, Rejecter = 5%	[15]
Probability of catch-up	55%	
Prob. administered first dose	Acceptor = 97%, Hesitant = 30%, Rejecter = 5%	
Prob. administered second dose	Acceptor = 98%, Hesitant = 82%, Rejecter = 33%	
Primary vaccine failure first dose	16–24%	[4, 12, 13]
Primary vaccine failure second dose	5–16%	[4, 12, 13]
Waning of vaccine immunity 1 dose	0.02 year ⁻¹	[13]
Waning of vaccine immunity 2 dose	0.00 year ⁻¹	[13]

PeerJ 6:e5012 (<https://doi.org/10.7717/peerj.5012>). CC BY 4.0 [40]

(PHN) in the statechart. A person afflicted by shingles will eventually recover, after which they are subject to a certain chance of relapsing.

Life and death are handled by another state chart in this model, shown in Fig. 9.8. A person is alive for the duration of their lifetime, with mortality occurring according to a certain background death rate or due to VZV infection (realized by receipt of a message for death sent from the infection process within the same agent). Certain events, including childbirth for females, happen during their life, as represented by the arrows within the alive state. The external self-transition updates information periodically while the person is alive. It also bears note that the model includes natality processes, and newborn babies enter the statechart through the initial transition, which extends down from the “statechartAging” label.

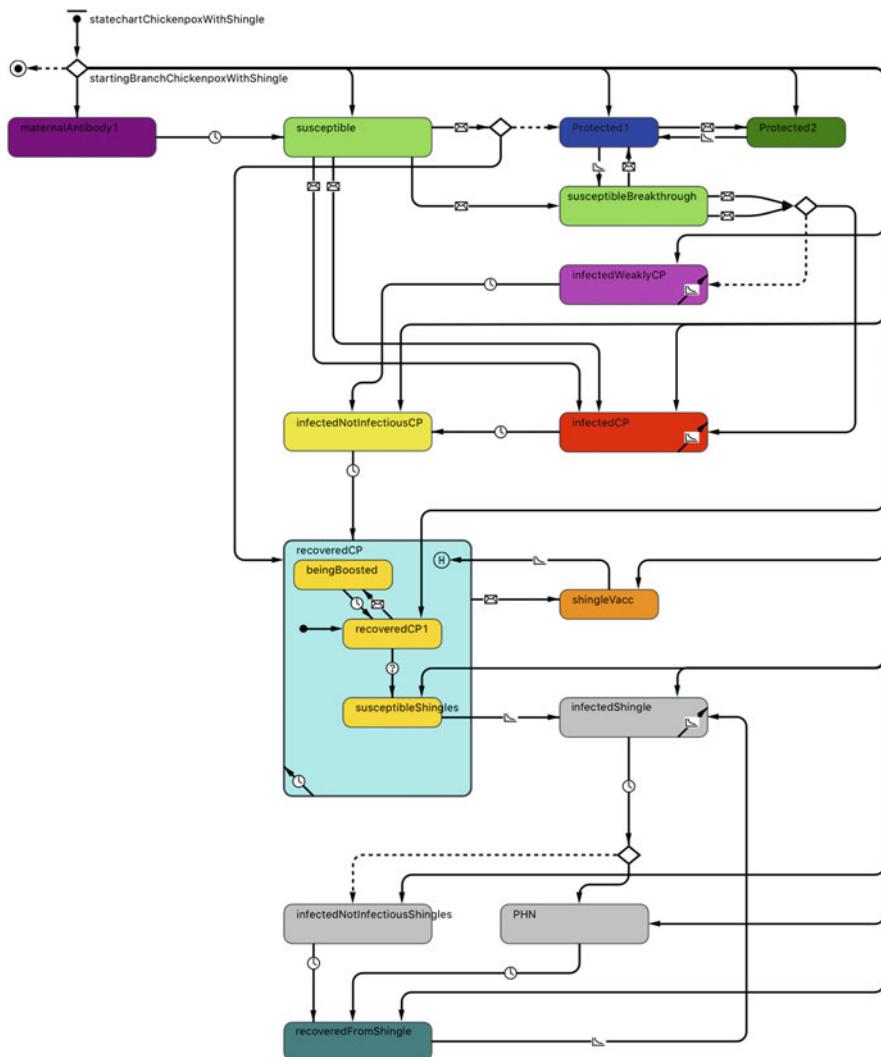
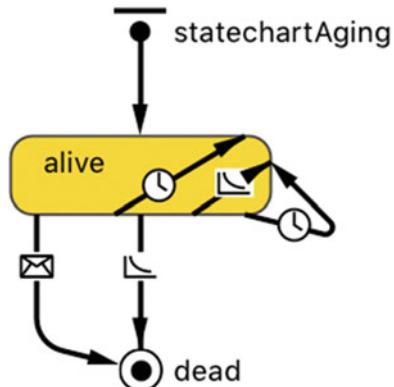


Fig. 9.7 Statechart for natural history of chickenpox and shingles. Image from Rafferty et al. (2018) Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling. *PeerJ* 6:e5012 (<https://doi.org/10.7717/peerj.5012>). CC BY 4.0 [40]

A further statechart—shown in Fig. 9.9—represents a person's adherence to the vaccination schedule. When a person is too young to be vaccinated, they have no scheduled vaccines. They will subsequently become due for their first dose and either receive it or not and eventually become due for a second dose and receive it or not. In the event they receive the second dose but missed the first, they may get a catch-up dose.

Fig. 9.8 Statechart for life and death in the chickenpox model. Image from Rafferty et al. (2018) Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling. *PeerJ* 6:e5012 (<https://doi.org/10.7717/peerj.5012>). CC BY 4.0 [40]



Heterogeneity is represented by these parallel statecharts; all of the statecharts described above were part of the person agent. Within a given agent, all such statecharts operate concurrently, with each describing different aspects of that person's state and actions by which that state evolves. Age is captured as a continuous quantity in this model, without the coarse-graining common in stratified aggregate models.

9.3.4 Model Fit

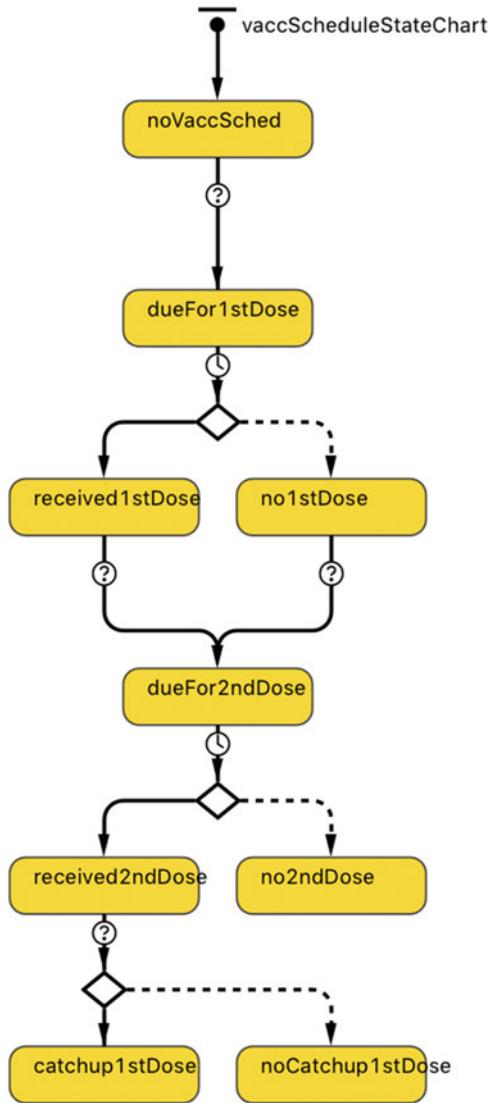
This model was fit to empirical data in several ways. First, we considered chickenpox incidence in the pre-vaccination era and its distribution over age groups, as seen in Fig. 9.10a. Red represents reference data from literature and blue represents the model output; the x-axis is age in years, and the y-axis is chickenpox incidence per 100,000 population for individuals of those ages.

A similar comparison was undertaken for shingles incidence, in Fig. 9.10b; again, red is the reference and blue is the model output. The flaring of the blue at the high age point is due to progressively smaller counts of people of in oldest ages.

Answering the question that was posed initially, the emergent behavior that we observed in this model suggests that chickenpox vaccination is expected to cause an increase in shingles cases before it leads to an eventual decrease. Referring to the plot in Fig. 9.11, along the x-axis is time (measured in years relative to the completion of the burn-in period), and the y-axis is total shingles cases in the model, with chickenpox vaccination beginning at time 0. The figure represents one pair of realizations, consisting of a baseline and an corresponding intervention. Once vaccination is introduced, the paths diverge, and there is an increase in shingles cases for a period of about 30–35 years, followed by a steep decrease.

The degree to which VZV exposure boosts immunity remains uncertain clinically [32], so we conducted a sensitivity analysis on this model by examining the outcomes of assuming different durations of boosting. Figure 9.12 shows those

Fig. 9.9 Statechart for vaccination in the chickenpox model. Image from Rafferty et al. (2018) Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling. *PeerJ* 6:e5012 (<https://doi.org/10.7717/peerj.5012>). CC BY 4.0 [40]



outcomes for six alternative durations. The panel in the upper left is the result from assuming the briefest duration of the boosting effect; the assumed durations progressively increase between panels from left to right and then top to bottom, with the bottom-right panel depicting a situation where we have assumed the longest duration of boosting. Such results reveal another emergent behavior, which is that the degree to which shingles cases are expected to increase following the introduction of vaccination—and the duration of time over which the number of such cases exceeds that expected absent vaccination—depends on assumptions about exposure-induced boosting of immunity. The mechanism is a cohort effect

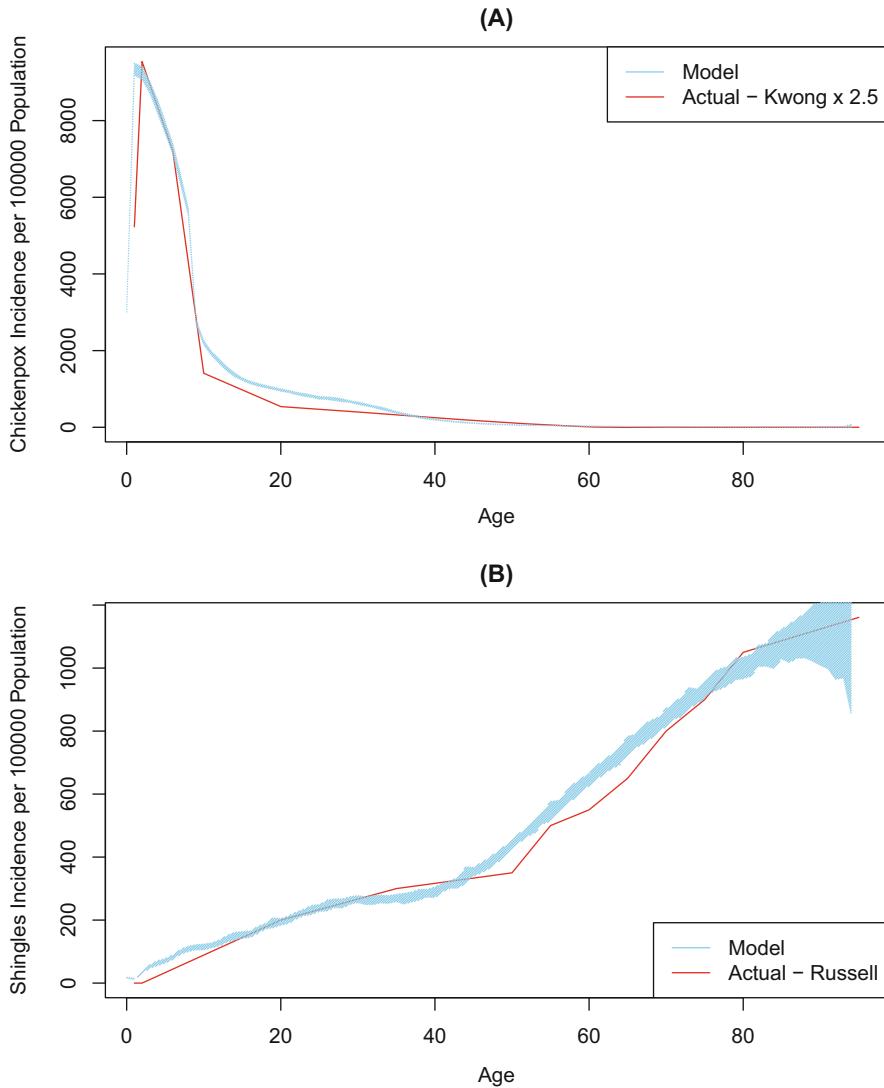


Fig. 9.10 Model fit for chickenpox and shingles incidence. Image from Rafferty et al. (2018) Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling. *PeerJ* 6:e5012 (<https://doi.org/10.7717/peerj.5012>). CC BY 4.0 [40]

where those infected with VZV immediately prior to the implementation of a vaccination program are denied the exposure-induced boosting that the older cohorts received due to low extant circulation of VZV and thus are more prone to developing shingles than the older cohorts, who were infected with the natural disease and benefited from ongoing boosting due to higher VZV circulation, as well as than the vaccinated cohorts.

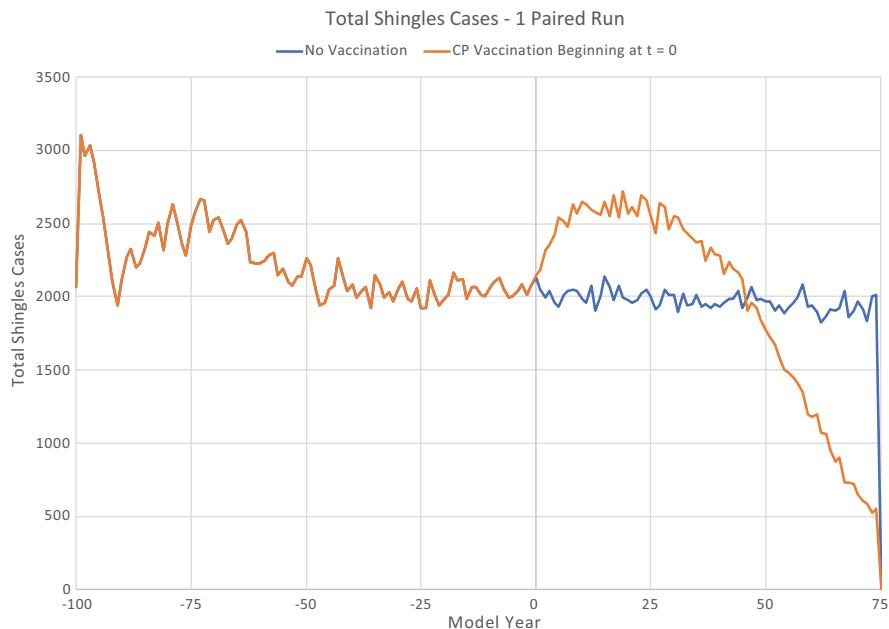


Fig. 9.11 Model result for a single paired run

9.3.5 Costs and QALYs

This model was further used to analyze some health economic questions; to do this, some assumptions regarding costs had to be introduced exogenously. These included costs of vaccine doses, costs for general practitioner and emergency department visits due to VZV, per-day costs for people who are hospitalized, personal expenses, and productivity loss costs.

Additionally, QALYs—which stands for quality-adjusted life years and is a common measure in health economics [57]—were accumulated in the model based on a person’s accumulated time in various states.

9.3.6 Suitability of ABM

The questions investigated were the following: (1) assessing the impact of chickenpox vaccination on shingles incidence [40]; (2) identifying an optimum, in terms of quality of life measures, vaccination schedule for chickenpox within Canada [42] (serving as an example of a longitudinal intervention); and (3) evaluating the cost-effectiveness of chickenpox vaccination with discounting [41]. These analyses play to the strength of ABMs in that they would be far more cumbersome, or

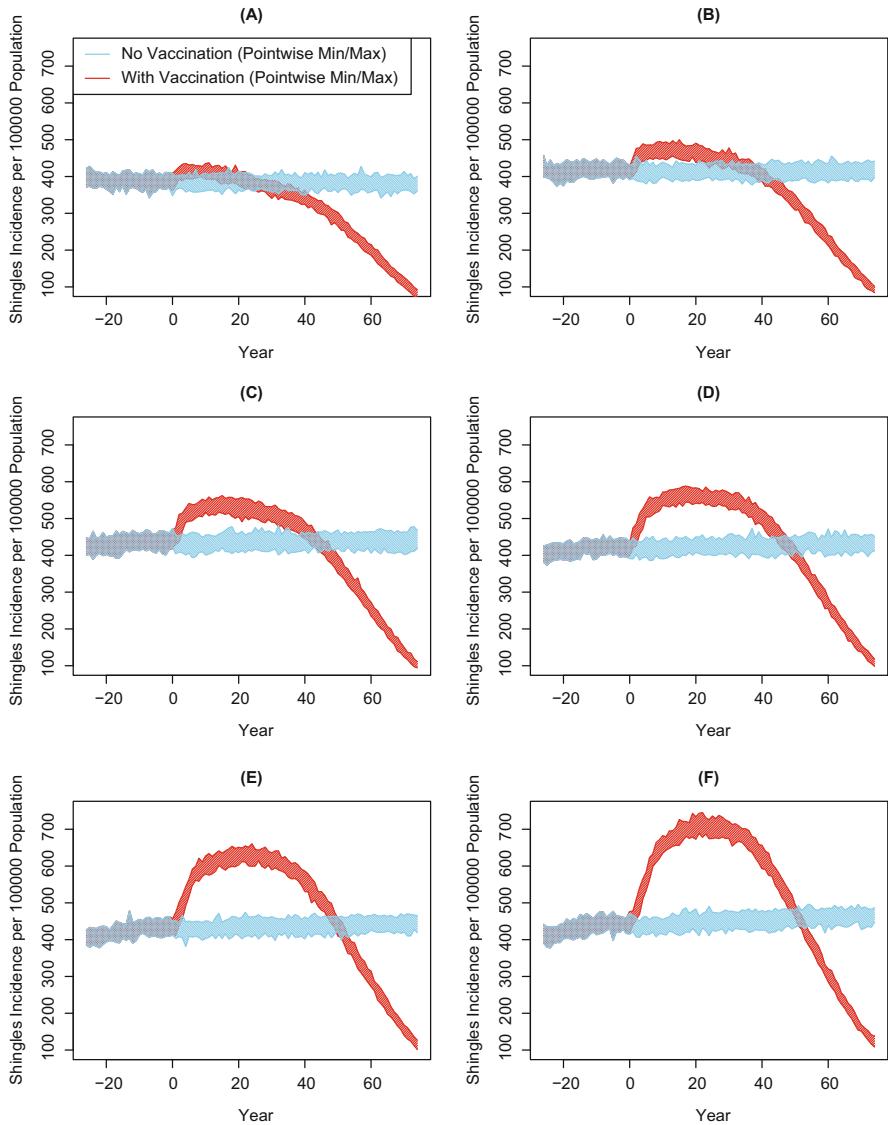


Fig. 9.12 Chickenpox model output varying duration of boosting: 2 years (a), 3 years (b), 4 years (c), 5 years (d), 6 years (e), and 7 years (f). Image from Rafferty et al. (2018) Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling. *PeerJ* 6:e5012 (<https://doi.org/10.7717/peerj.5012>). CC BY 4.0 [40]

impossible, to investigate using an equally fine-grained lens with compartmental modeling techniques.

It is worth pointing out that other modeling studies [5, 38] have approached similar research questions using compartmental modeling techniques. In comparison with the example presented here, these works made strict assumptions about the effect of boosting and, necessarily, employed coarse-grained age groups and were limited in their ability to examine the relationship between individual vaccination and exposure history and health outcomes.

9.3.7 *Choice of AnyLogic as a Tool*

While there are many software packages that facilitate the programming of ABMs, the authors feel that AnyLogic is preferable for some of their work due to its ability to combine agent-based, system dynamics, and discrete event simulation logic in the construction of hybrid models. While this strength is not brought to bear in the examples presented, which are purely agent-based, this is a key factor in motivating author selection of AnyLogic as the tool for a number of projects. A further central motivator for this project was the capacity of AnyLogic’s declarative modeling language to communicate model assumptions and logic to health scientists on the modeling team lacking computational training, and to allow such scientists to directly critique, refine, and manipulate and modify model assumptions and scenarios. Extending accessibility and transparency of the large majority of model assumptions across the entire interdisciplinary team can greatly reduce risk of misunderstandings, miscommunications, and resulting model design and implementation errors. By facilitating more effective team science, such transparency materially elevates the team’s ability to produce rigorous, relevant, and impactful models.

9.4 Example: Pertussis

The second example to be discussed is a simulation model of pertussis [20] which was also developed using the agent-based methods with AnyLogic software.

9.4.1 *Pertussis*

Pertussis, commonly known as whooping cough, is a respiratory infection caused by the *Bordetella pertussis* bacterium and transmitted by droplets in the air. Symptomatic individuals develop a characteristic whooping sound—from which the common name is derived—as they gasp for breath after extended bouts of coughing. Infants are at high risk of serious and even lethal complications; two-thirds experience trouble breathing, and half are hospitalized. Adults, by contrast,

are often asymptomatic or have nonspecific symptoms, which may be confused with a cold or flu. Risk of complications rises with age, smoking, and pre-existing asthma or other respiratory conditions. A high rate of waning of immunity for this disease necessitates a course of six to eight vaccinations to achieve full protection; many people don't complete this entire course of vaccination [10, 17, 53, 59].

In Canada, the pre-vaccination era incidence of pertussis was characterized by multiyear cycles, infecting mostly children. Upon widespread vaccination, pertussis incidence plummeted. The whole-cell vaccine was the first vaccine developed, and it achieved high efficacy but suffered from a higher risk of side effects compared to later-developed vaccines. There were issues, beginning as early as the 1970s and 1980s, with misinformation instigating vaccine hesitancy and impairing vaccine coverage [17, 49].

An acellular vaccine was introduced in the 1990s, and later the multivalent DTaP vaccine, which reduced side effects, at the cost of lower efficacy. Vaccination complacency and hesitancy caused vaccination rates to flag, and Canadian outbreaks began in earnest in the 2010s, concentrating in under-vaccinated adolescents. Of particular concern was the growing risk that such transmission imposed on infants, who constitute the primary risk group [17, 49].

In Alberta, which was the focus of our study, vaccination rates for doses 4 and later are in the 70–80% range overall. Critically, however, there are notable disparities in vaccination rates between families, schools, and communities, leading to poorly vaccinated geographic areas and regions of contact networks at high risk of outbreaks. Notable outbreaks have occurred in the last 10–15 years, many in under-vaccinated adolescents who are clustered in communities with a high density of vaccine-hesitant or vaccine-refusing individuals [25].

9.4.2 *Model Scope*

Before characterizing model structure, a few words about model scope are in order. Endogenously represented factors within the model included transmission of pertussis, contact patterns, schools and school transitions, household structure, household formation once a person reaches adulthood, immune protection—represented as active and passive protection on a continuous-state basis—family vaccination adherence, vaccine catch-up for those who miss doses but remain accepting of vaccines, fertility, mortality, differential case reporting based on infection severity, vaccine effectiveness, and the population preventable fraction, a measure related to vaccine effectiveness. Exogenously specified elements included vaccine attitudes and demographics of the initial population. Hyperparameters for some distributions for other parameters—including vaccine attitude, immune memory types and waning rates where whole-cell, acellular, and natural infection-induced immunity—were represented distinctly. School count and characteristics and age-specific ascertainment rates for more serious infections were also specified exogenously. Ignored were household type change, interregional mixing, pertussis

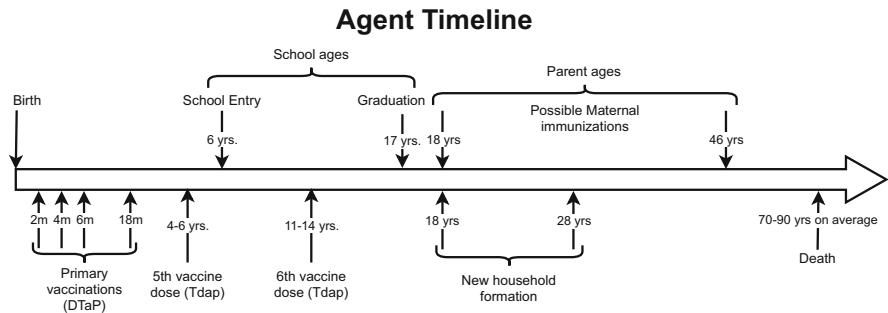


Fig. 9.13 Agent timeline for pertussis model. [20]

hospitalizations, and mortality. Other than contacts in homes and schools, the model also ignored occurrence of contacts during the day, including those at workplaces and childcare venues.

9.4.3 Model Structure

Figure 9.13 represents the timeline of an agent over its life course, beginning with birth. Vaccination occurs at various points mostly during childhood; children enter school at 6 years and complete at 18 years; between the ages of 18 and 28, children move out and form their own households and can become parents on their own up to 46 years. On average, 70–90 years is the life expectancy.

Pregnancy and fertility are represented with an age- and family-structure-specific fertility rate. The statechart in Fig. 9.14 represents states related to pregnancy, where a person begins in a nonpregnant state. The occurrence of conception is characterized by a (hazard, i.e., temporal probability density) rate dependent on that person's fertility rate, as given by their age and the number of previous children that they've had. Following conception, that person progresses through the trimesters and ends in a postpartum period.

Vaccine scheduling and compliance involved people being either *on schedule* or *noncompliant* and subject to a hazard rate of switching between those states at rates dependent on their vaccine attitude, as represented in the statechart shown in Fig. 9.15. Each person agent in the model is randomly assigned a number representing vaccine acceptance, where 1 represents full acceptance and 0 full refusal. Multi-person households act according to the minimum vaccine acceptance of the parents. Distributions were calibrated to ensure that the emergent vaccine coverage from the model matched the empirical data.

Within the ABM, contact patterns constitute an emergent property of the model. Figure 9.16 depicts an example of such contact network for a given person, represented by the yellow dot in this illustration; from the picture, it can be readily

Fig. 9.14 Fertility statechart in the pertussis model

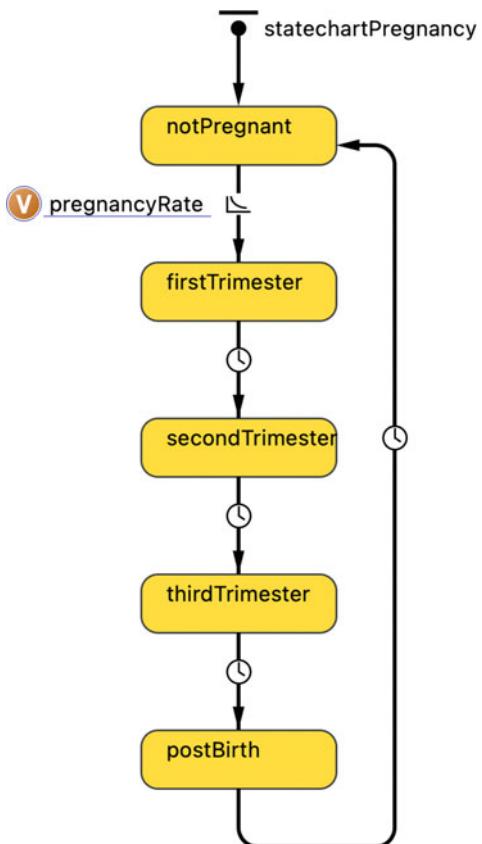
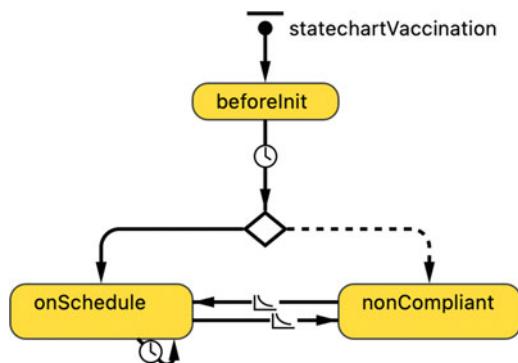


Fig. 9.15 Vaccine compliance statechart in the pertussis model



discerned that the network is composed of several distinct types of connections. That index person shown in yellow is connected to all other people within a certain radius according to background contacts indicated in red. If that person is a child, they would be connected to all other children at their school. Every person is associated with a household and is connected to all other people within their own household.

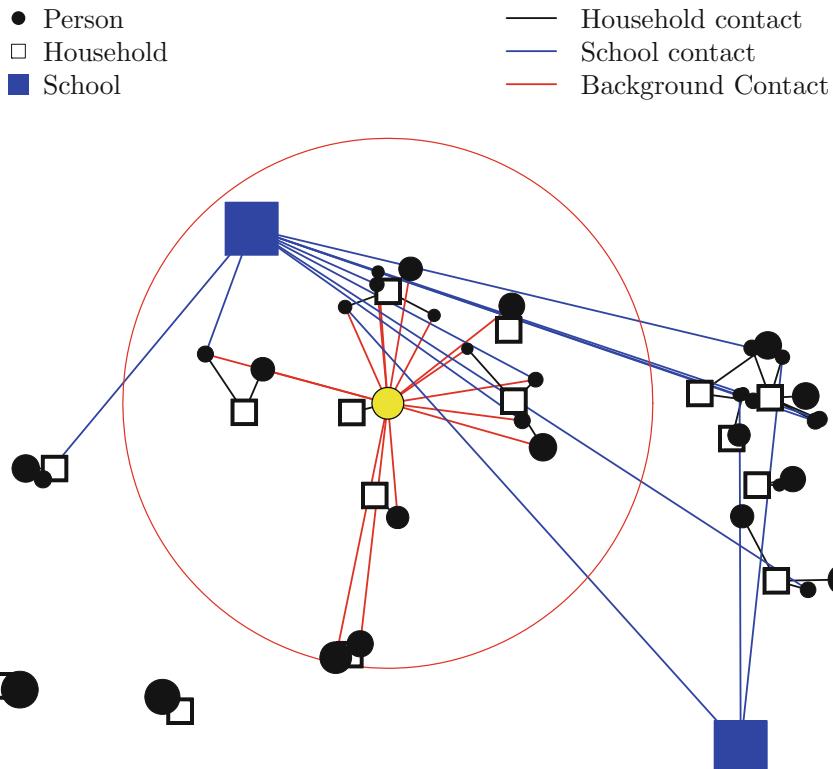


Fig. 9.16 Contact network in the pertussis model [20]

Immunity is represented as a continuous quantity, between 0 and 1, for each person. Figure 9.17 represents the development of active protection from the first five doses of the vaccine course, following the Alberta schedule for pertussis vaccination. A person begins immune naive; progression in the vertical direction is triggered by administration of a vaccine dose; between doses and absent occurrence of infection, immunity wanes exponentially.

Protection level was determined by Eq. 9.1, where maternal immunity, called passive protection, decays at a rapid rate, while active protection is determined by immune memory and decays according to a rate specific to the supporting immune memory type: natural disease, whole-cell vaccine, or acellular vaccine.

$$p = \min(p_{\text{active}} + p_{\text{passive}}, 1.0) \quad (9.1)$$

Figure 9.18 represents the effect of maternal vaccination. Solid black represents the mother's immunity *without* maternal vaccination, and the dotted black line represents the child's immunity, which decays rapidly following birth. Solid blue represents the effect of a mother receiving a vaccination during the third trimester

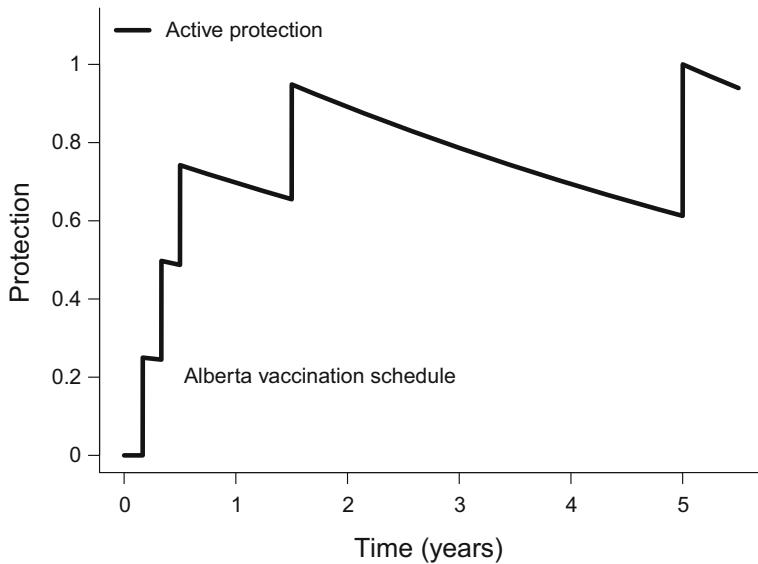


Fig. 9.17 Active protection generated by the first five doses of the vaccine course [20]

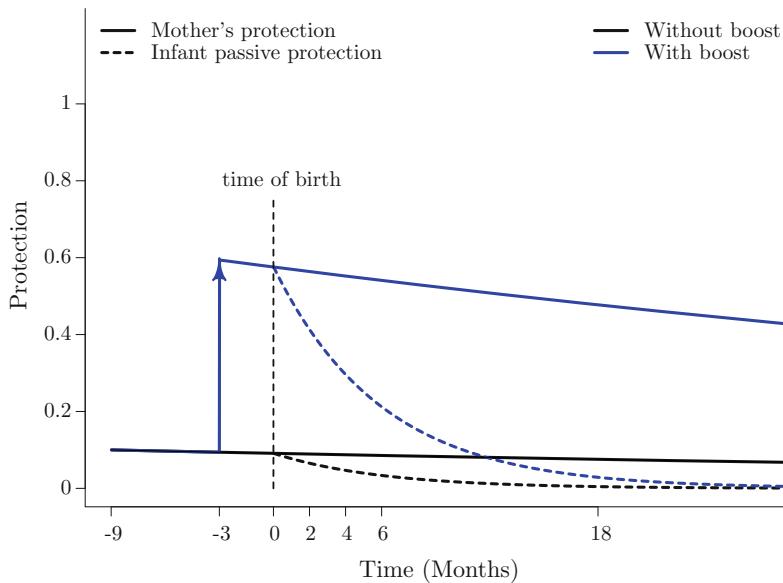


Fig. 9.18 Effect of maternal immunization on infant immunity [20]

of pregnancy (depicted as time—3 months), and dotted blue line represents the infant under this scenario. Identical decay immunity rates obtain in both the vaccination and non-vaccination cases, but the occurrence of vaccination allows immune memory in both mother and child to decay from a higher point, leaving a higher protection level in the child for the first year of life.

A person in the model is exposed to pertussis with a certain per-exposure probability of infection π , carried in a message passed from an infective to a susceptible person. An exposed susceptible is infected with a probability π if their protection level lies below some threshold, α_i . Following infection, active protection is boosted to a maximum level and subsequently begins to decay. Vaccination leads to active protection being boosted incrementally as shown in Fig. 9.17. For infants, the blunting hypothesis [22] suggests that maternal immunization may lead to impaired immunity development by the child when receiving early doses of the vaccine.

9.4.4 Model Fit

Demographics were informed by the census population pyramid, the initial distribution of household types, and distributions of couple and single households by the number of children in the household [16, 51].

Vaccine coverage is endogenous to the model, and it was fit to actual coverage in Alberta, as shown in Fig. 9.19, which shows dose along the x-axis and vaccine coverage on the y-axis with the gray columns representing the true state in Alberta and the green columns representing what the model achieved.

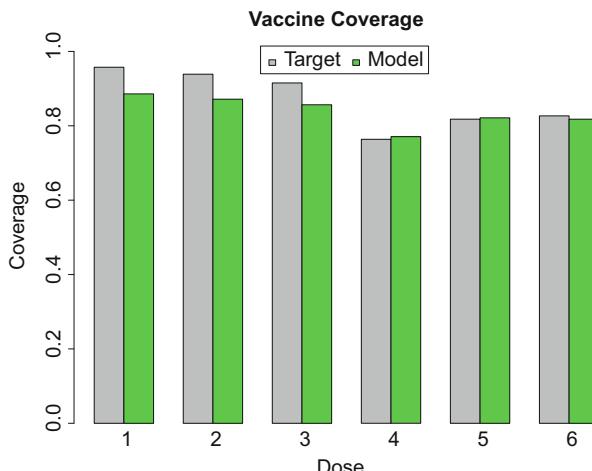


Fig. 9.19 Vaccine coverage in the pertussis model [20]

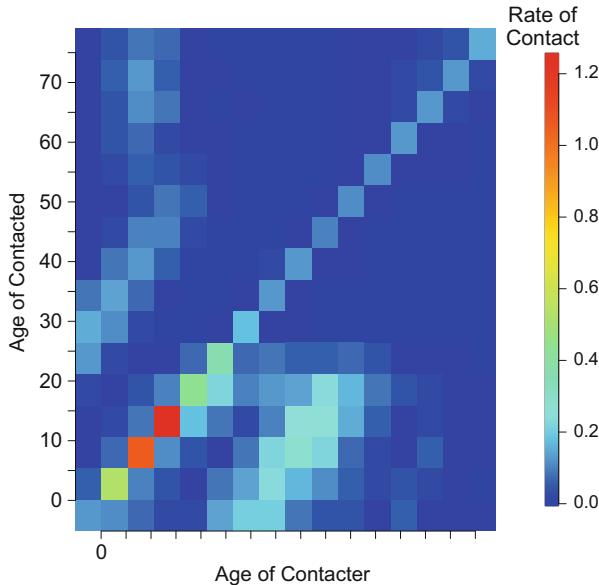


Fig. 9.20 Emergent contact patterns from the pertussis model [20]

Figure 9.20 represents the emergent contact patterns resulting from the contact structure described in Fig. 9.16. The diagonal represents people contacting other people of approximately their own age, the off-diagonal areas represent child-parent contacts, and then some of the other outlying areas represent child-grandparent contacts. This was compared to Mossong et al. [30], also known as POLYMOD, done in Europe.

Figure 9.21 represents calibration and validation matches; vaccine coverage was discussed above, and we also matched to the average risk ratio since the last vaccination dose, mean yearly incidence, density of yearly incidence, autocorrelation of yearly incidence, and age distribution of pertussis incidence. The goal of using these measures to fit the model was to ensure that outbreaks in the model were of size and frequency that are expected based on empirical data without trying to force it to follow every peak and valley of the actual historical outbreak pattern.

9.4.5 Scenarios

All scenarios involved running the model for a time horizon of 50 years with a 20-year burn-in, or warm-up, period before maternal immunization began. The initial population was 500,000 but, with an open population involving births and deaths endogenous to the model, that population expanded and shrank over time. The baseline involved no vaccination. The main intervention was maternal immunization

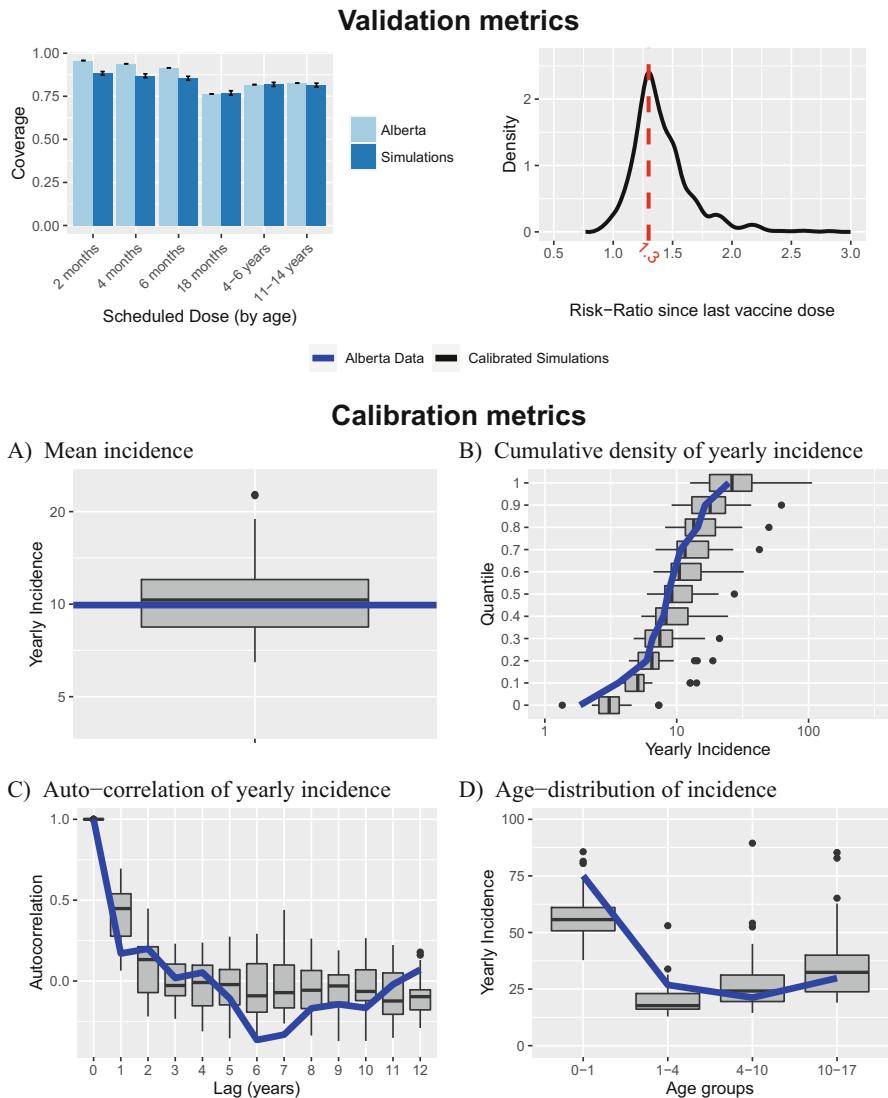


Fig. 9.21 Pertussis model fit [20]

administered across 50% of pregnancies. A number of sensitivity analyses were examined, including different rates of maternal vaccination, alternative durations of maternal-antibody-based immunity, varying ascertainment rates, blunting, and no passive protection from immunization. All scenarios were assessed in ensembles of 30 realizations.

Model outcomes suggest that immunization is highly effective in reducing infant infections and substantial benefits are provided by both cocooning and transfer of passive immunity.

9.4.6 *Suitability of ABM*

A number of features of this study are well-suited to be examined with an ABM. One is the need to consider network effects, particularly in light of an overall contact network emergent from a number of time-varying subnetworks. Continuous, heterogeneous state in the form of dynamics of immune memory of a person is readily captured using an ABM, but can only be discretely approximated using a stratified compartment model. As a family-level phenomenon, the adequate characterization of cocooning-based protection is difficult to achieve in a compartmental model, but readily with an ABM. While readily characterized with an ABM, description of targeted intervention of immunization at a certain point during pregnancy would be highly cumbersome within a compartmental model due to the curse of stratification dimensionality. This model can also be held up as an exemplar of large compute requirements as a limitation of ABM, with a single realization taking over a day to evaluate on a dedicated computer platform.

9.5 Trade-Offs Between ABMs and Aggregate Models

This section briefly highlights some of the many trade-offs between agent-based and aggregate models, including those based on both technical considerations and those associated with the modeling process and organizational context. Those seeking further depth of discussion of such trade-offs are referred to past contributions focusing on this subject [2, 26, 34–36, 39, 43, 45].

An ABM advantage of great significance in both conceptual and practical domains lies in its ability to readily capture and nimbly evolve continuous, discrete, and relational heterogeneity, whether static or evolving over time. Straightforward and easily evolved representation of heterogeneity allows us to more readily address concerns involving health equity and disparities. We have multiple aspects of state we can readily keep track of, for example, multiple comorbid conditions or behavioral concerns.

While it represents a type of heterogeneity, the ability to capture arbitrary aspects of individual history itself is of sufficient import to merit further remark. Longitudinal information from the model can be compared against comparable empirical longitudinal information—information that is widespread and increasingly commonly available. We can calibrate the model against such longitudinal information. As will be further noted below, such longitudinal information can play

a key role in supporting interventions. Moreover, such longitudinal information is frequently a very important class of data for understanding dynamics.

Also of particular value to the modeling process and learning with a model is the ability to nimbly evolve the representation of heterogeneity, quickly adding new dimensions of dimensionality to the model, rapidly altering the representation of certain types of heterogeneity according to learning regarding the important distinction and measures, or removing it when judged appropriate.

Individual-based models, including ABMs, are better for examining fine-grained consequences involving network and space effects. Using multi-scale modeling, we can represent multilevel nesting of such context, representing—in accordance with the socio-ecological model [6–8]—the successively broad nesting of a person within a neighborhood, a school, a municipality, and a country, for example. Capturing such layers of context is also advantageous to examine emergent behavior across such different levels of scale. A further advantage concerns ABM’s very natural means of representing such nested contexts: in contrast to the “horizontal” character of compartmental models, nesting within ABMs mirrors, in a very natural way, nesting in the world.

Although ABMs can readily handle data or data gaps that require imposing homogeneous assumptions across population members, such contexts eliminate one of the important competitive advantages of agent-based modeling. However, the ability to scalably and nimbly characterize heterogeneity in individual characteristics is just one motivator for use of ABMs; an agent-based approach may still be recommended by its other strengths, such as the ability to capture agent history, multidimensional state evolution, and network and spatial effects as such factors bear on model governing processes, intervention mechanisms, or outcomes of interest.

ABM’s representation of agents as situated in contexts further allows capturing of situated perception of individuals, learning over time, and decision-making given those perceptions.

Agent-based modeling’s finer resolution supports characterizing and examining interventions at a far more detailed level than is readily possible with aggregate modeling. While aggregate models can often be used to secure answers to coarse-grained questions as to *where* to intervene in a system, ABMs can go further by examining *how* to best intervene [28]. Within the sphere of interventions, ABMs can also be used to examine matters involving the *implementation* of interventions, dealing with the sphere of implementation science by evaluating intervention scalability, rollout and scale-up dynamics, financial sustainability, and the time to effect. Also, because of ABM’s ability to characterize individual progression and layers of context, we can represent interventions that are contextualized and that are based on individual position in networks or space, and we can examine interventions that are highly targeted in ways that really are not readily addressable in any plausible fashion with aggregate models. The ability of ABMs to information on arbitrary aspects of individual history capture is of such importance for intervention assessment as to merit further discussion. Such longitudinal data readily supports ABM investigation of the broad and widespread class of interventions that target

or trigger interaction based on history at an individual level or at an intermediate level of scale (e.g., at a family or neighborhood level). Such interventions are both common and important and can readily involve targeting or triggering rules that are difficult to adequately characterize using aggregate models; the contrasting ease with which such interventions can be represented in ABMs renders them of great value in many public health contexts.

Of final note is ABM's capacity to support critical evaluation of broad classes of data collection and analysis methods by virtue of ABMs to characterize such methods against synthetic data in an *in silico* environment in which the underlying situation (the “ground truth”) is in fact known—an approach commonly referred to as “simulation experiments” within statistics. Through such ABM-based evaluation, it is possible to more proactively identify blind spots within such data collection and analysis methods. For example, [37] sought to use machine learning strategies such as particle filtering to assess on a recurrent basis the underlying epidemiological state of some contexts. A textured agent-based model provided a ready means of evaluating the effectiveness of such methods by serving as a source of synthetic empirical data and comparing the particle filter-based estimates of the underlying state against the “synthetic ground truth” given by the actual situation within the agent-based model. The same experimental setup can readily allow for studying under what conditions the estimates are more—or less—accurate, the impact of network type, or frequency of data collection on estimate accuracy [48]. More broadly, similar methods can be easily used—*mutatis mutandis*—to assess the accuracy of other sampling methods, inference strategies, and adequacy of associated data collection mechanisms.

With models, it is often valuable to engage in storytelling when engaging with stakeholders—whether people with lived experience, policymakers, or other knowledge users. Models are exceptionally powerful as storytelling vehicles when we can link them up to the experiences of individuals and organizers. ABMs, in particular, can excel at this by showing or recounting as narrative one or more simulation trajectories of individual agents (including aspects of history) or of different components of an organization [27].

All techniques have limitations, and ABMs are no exception. Explainability to non-modelers is currently a foremost challenge. While visual depiction of dynamic models is a key asset in securing stakeholder feedback regarding and critique of such models, in the current state of the art, there is no unifying visual description language for depicting ABM structure. Moreover, there is no widespread (much less universal) mathematical framework in which such models are specified. Instead, the structure and rules underlying a model are operationally specified in code—code that is almost always inaccessible to stakeholders elsewhere on model teams. Even small ABMs commonly require a modest amount of code; medium-sized production ABMs are commonly accompanied by sizable codebases. Beyond impairing transparency to and critique by stakeholders, the lack of formal, transparent model specification and the frequently sprawling nature of ABM codebases pose notable problems for the communication and replication of model results that lie at the basis of scientific advances.

With all types of dynamic modeling seeking to address questions and characterize important types of factors within the world, the basic issue of model validation is: “have I built the right model?” But with ABMs, the fundamental question of model verification—“have I built the model right?”—achieves particular texture, importance, and operational urgency. This particularly reflects the fact that because of the amount of software engineering they require during the model implementation stage, ABMs often contain a great deal of programming logic where a significant number of bugs may lurk. Building and maintaining medium-sized ABMs requires not only the traditional interdisciplinary mix essential for supporting other impactful dynamic modeling projects but also solid software engineering skills. Such efforts place a premium on practice of quality assurance skills such as pair modeling, peer desk checks, and formal model inspections [55, 58], model testing and mocking [55], and continuous integration [55]. They also require much effort by modelers to avoid the risk of “not being able to see the forest on account of the trees”—being so distracted by the welter of implementation-level software engineering detail that they lose clarity regarding and reasoning about model structure. Finally, the large volumes of code require interplay of skilled software engineering and savvy modeling to ensure that a model can remain capable of evolving nimbly with learning.

A key shortcoming of agent-based modeling is also the flip side of one of its key strengths: flexibility. While that flexibility offers great advantages in crafting models that offer high-resolution lenses to investigate public health questions, it’s all too easy to take the flexibility of ABMs and run afoul of it by building models where too much is included. When building such models, it is key to apply the YAGNI principle (“you ain’t gonna need it”) [21] and building it up in an agile fashion bit by bit [1, 3, 23, 55].

It’s fair to say that aggregate models frequently have an edge in terms of faster (albeit more abstract) construction, lower computational burden, and greater transparency. Because of the ability to represent aggregate models as ordinary differential equations or stochastic differential equations, they can be analyzed formally and mathematically understood in ways that are often more immediate than what is possible with ABMs. Aggregate models have lower baseline cost and involve far less programming than ABMs. In terms of run (numerical integration) time, aggregate models’ computational performance costs are invariant to the population size. The lack of stochastics in ordinary differential equation and other deterministic compartmental models means that you can run the model quite directly without as much need for ensembles. Overall, the fact that you can build aggregate models more quickly and run them more quickly leaves more time for learning and refinement. While we can say that aggregate models are often simpler, some mechanisms can be simpler to describe in ABMs, such as those many points of understanding or theory characterizing phenomena at or benefiting from description at an individual level. Representing multiple aspects of heterogeneity—static or dynamic—in aggregate compartmental model gives rise to a combinatorial explosion of structure. With many ABM packages, we can readily use visualization of model outputs to aid communication and intuition. However, because of the

large amounts of code involved in contemporary agent-based modeling practice, ABM structure is frequently considerably less accessible and transparent to project stakeholders compared to what is possible in compartmental modeling.

9.6 Summary

Agent-based modeling, in summary, is a powerful tool for investigating health-related questions, allowing us to represent individual history and targeted interventions while capturing supportive spatial, geographic, or network context. ABMs can capture agent-environment interactions in rich ways with GIS and irregular spatial networks. ABMs richly capture heterogeneity, particularly the ability to capture heterogeneity of individual history impacts, early life insults, or adverse childhood experiences, which are key for addressing specific health equity needs. Key limitations of agent-based models include computational expense: a single realization can require hours to run and requirements scale up with population; this is exacerbated by the stochastic nature of ABMs, which necessitates running an ensemble of realizations to fully capture the regularities of the model. The lack of a crisp mathematical description or visual language for ABMs impairs modeling transparency to stakeholders and communication of modeling results. For all of their trade-offs, it is important to recognize that recent modeling advances point us to look beyond choosing one or the other modeling method, and to the importance of judiciously weaving them together for effective hybrid modeling. We defer such discussions of hybrid modeling and exciting advances toward declarative modeling to later contributions.

Acknowledgments The authors would like to acknowledge our collaborators on these works, Drs. Alexander Doroshenko, Karsten Hempel, Weicheng “Winchell” Qian, and Ellen Rafferty, as well as the Canadian Immunization Research Network (CIRN) and Mathematics for Public Health (MfPH). Co-author Osgood wishes to express his appreciation of support from NSERC via the Discovery Grants program (RGPIN 2017-04647) and from SYK & XZO.

References

1. Abrahamsson, P., Salo, O., Ronkainen, J., Warsta, J.: Agile software development methods: review and analysis (2017). arXiv preprint arXiv:1709.08439
2. Bankes, S.C.: Agent-based modeling: a revolution? Proc. Natl. Acad. Sci. **99**(suppl 3), 7199–7200 (2002)
3. Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., et al.: Manifesto for agile software development (2001)
4. Bonanni, P., Gershon, A., Gershon, M., Kulcsár, A., Papaevangelou, V., Rentier, B., Sadzot-Delvaux, C., Usonis, V., Vesikari, T., Weil-Olivier, C., de Winter, P., Wutzler, P.: Primary versus secondary failure after varicella vaccination. Pediatr. Infect. Dis. J. **32**(7), e305–e313 (2013)

5. Brisson, M., Melkonyan, G., Drolet, M., Serres, G.D., Thibeault, R., Wals, P.D.: Modeling the impact of one- and two-dose varicella vaccination on the epidemiology of varicella and zoster. *Vaccine* **28**(19), 3385–3397 (2010)
6. Bronfenbrenner, U.: Toward an experimental ecology of human development. *Am. Psychologist* **32**(7), 513 (1977)
7. Bronfenbrenner, U.: Ecology of the family as a context for human development: research perspectives. *Dev. Psychol.* **22**(6), 723 (1986)
8. Bronfenbrenner, U.: Ecological Systems Theory. Jessica Kingsley Publishers (1992)
9. Campbell, A., Ismail, S., Tan, B.: Literature review on one-dose and two-dose varicella vaccination. *Can. Communicable Dis. Rep.* **36**(ACS-10), 1–24 (2010). <https://doi.org/10.14745/ccdr.v36i00a10>
10. Centers for Disease Control and Prevention: Pertussis frequently asked questions (2019). <https://www.cdc.gov/pertussis/about/faqs.html>
11. Cohen, J.I.: Herpes zoster. *N. Eng. J. Med.* **369**(3), 255–263 (2013). <https://doi.org/10.1056/nejmcp1302674>
12. Duncan, J.R., Witkop, C.T., Webber, B.J., Costello, A.A.: Varicella seroepidemiology in united states air force recruits: a retrospective cohort study comparing immunogenicity of varicella vaccination and natural infection. *Vaccine* **35**(18), 2351–2357 (2017)
13. Gershon, A., Takahashi, M., Seward, J.: Varicella Vaccine, pp. 837–869. Elsevier Saunders, Philadelphia (2012)
14. Gold, L., Shiell, A., Hawe, P., Riley, T., Rankin, B., Smithers, P.: The costs of a community-based intervention to promote maternal health. *Health Educ. Res.* **22**(5), 648–657 (2006). <https://doi.org/10.1093/her/cyl127>
15. Government of Alberta: Interactive Health Data Application. http://www.ahw.gov.ab.ca/IHDA_Retrieval/
16. Government of Alberta: Open government program (2019). <https://open.alberta.ca/opendata>
17. Government of Canada: Pertussis (whooping cough) (2019). <https://www.canada.ca/en/public-health/services/immunization/vaccine-preventable-diseases/pertussis-whooping-cough.html>
18. Hammond, R.A.: Peer reviewed: complex systems modeling for obesity research. *Preventing Chron. Dis.* **6**(3) (2009)
19. Hawe, P., Shiell, A., Riley, T., Gold, L.: Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. *J. Epidemiol. Community Health* **58**(9), 788–793 (2004)
20. Hempel, K., McDonald, W., Osgood, N.D., Fisman, D., Halperin, S.A., Crowcroft, N., Klein, N., Rohani, P., Doroshenko, A.: Evaluation of the effectiveness of maternal immunization against pertussis in Alberta using agent-based modeling: a Canadian Immunization Research Network study. *Vaccine* (Under review following resubmission for minor revisions)
21. Jeffries, R., Hendrickson, M.M., Anderson, A., Hendrickson, C.: Extreme programming installed. Addison-Wesley Professional (2001)
22. Kandil, W., Savic, M., Ceregido, M.A., Guignard, A., Kuznetsova, A., Mukherjee, P.: Immune interference (blunting) in the context of maternal immunization with TDAP-containing vaccines: is it a class effect? *Expert Rev. Vaccines* **19**(4), 341–352 (2020). <https://doi.org/10.1080/14760584.2020.1749597>
23. Krueger, L.K., Qian, W., Osgood, N., Choi, K.: Agile design meets hybrid models: Using modularity to enhance hybrid model design and use. In: 2016 Winter Simulation Conference (WSC), pp. 1428–1438. IEEE (2016). <https://doi.org/10.1109/WSC.2016.7822195>
24. Kwong, J., Tanuseputro, P., Zagorski, B., Moineddin, R., Chan, K.: Impact of varicella vaccination on health care outcomes in Ontario, Canada: effect of a publicly funded program? *Vaccine* **26**(47), 6006–6012 (2008)
25. Liu, X.C., Bell, C.A., Simmonds, K.A., Svenson, L.W., Fathima, S., Drews, S.J., Schopflocher, D.P., Russell, M.L.: Epidemiology of pertussis in Alberta, Canada 2004–2015. *BMC Public Health* **17**(1), 539 (2017)
26. Macal, C.M., North, M.J.: Agent-based modeling and simulation. In: Proceedings of the 2009 Winter Simulation Conference (WSC), pp. 86–98. IEEE (2009)

27. McDonald, G.W., Bradford, L., Neapetung, M., Osgood, N.D., Strickert, G., Waldner, C.L., Belcher, K., McLeod, L., Bharadwaj, L.: Case study of collaborative modeling in an indigenous community. *Water* **14**(17), 2601 (2022)
28. McDonnell, G.: Personal communication
29. Mejia Salazar, M.F., et al.: Social dynamics among mule deer and how they visit various environmental areas: implications for chronic wasting disease transmission. Ph.D. Thesis, University of Saskatchewan (2017)
30. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., et al.: Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**(3), e74 (2008)
31. Neumann, J.v.: Theory of Self-Reproducing Automata. Edited by Arthur W. Burks (1966)
32. Ogunjimi, B., Damme, P.V., Beutels, P.: Herpes zoster risk reduction through exposure to chickenpox patients: A systematic multidisciplinary review. *PLoS ONE* **8**(6), e66485 (2013). <https://doi.org/10.1371/journal.pone.0066485>
33. Ogunjimi, B., Willem, L., Beutels, P., Hens, N.: Integrating between-host transmission and within-host immunity to analyze the impact of varicella vaccination on zoster. *Elife* **4**, e07116 (2015)
34. Osgood, N.: Representing heterogeneity in complex feedback system modeling: computational resource and error scaling. In: 22nd International Conference of the System Dynamics Society (2004)
35. Osgood, N.: Using traditional and agent based toolsets for system dynamics: present tradeoffs and future evolution. In: Proceedings of the 25th International Conference of the System Dynamics Society, p. 19pp. Boston (2007)
36. Osgood, N.: Representing progression and interactions of comorbidities in aggregate and individual-based systems models. In: Proceedings of the 27th International Conference of the System Dynamics Society. Albuquerque, New Mexico (2009)
37. Osgood, N., Liu, J.: Towards closed loop modeling: Evaluating the prospects for creating recurrently regrounded aggregate simulation models using particle filtering. In: Proceedings of the 2014 Winter Simulation Conference, WSC '14, pp. 829–841. IEEE Press, Piscataway (2014)
38. Ouwend, M.J., Littlewood, K.J., Sauboin, C., Téhard, B., Denis, F., Boëlle, P.Y., Alain, S.: The impact of 2-dose routine measles, mumps, rubella, and varicella vaccination in France on the epidemiology of varicella and zoster using a dynamic model with an empirical contact matrix. *Clin. Ther.* **37**(4), 816–829.e10 (2015)
39. Parunak, H.V.D., Savit, R., Riolo, R.L.: Agent-based modeling vs. equation-based modeling: a case study and users' guide. In: Sichman, J.S., Conte, R., Gilbert, N. (eds.) Multi-Agent Systems and Agent-Based Simulation, vol. 1534, pp. 10–25. Springer, Berlin (1998). https://doi.org/10.1371/10.1007/10692956_2
40. Rafferty, E., McDonald, W., Qian, W., Osgood, N.D., Doroshenko, A.: Evaluation of the effect of chickenpox vaccination on shingles epidemiology using agent-based modeling. *PeerJ* **6**, e5012 (2018). <https://doi.org/10.7717/peerj.5012>
41. Rafferty, E.R., McDonald, W., Osgood, N.D., Doroshenko, A., Farag, M.: What we know now: an economic evaluation of chickenpox vaccination and dose timing using an agent-based model. *Value Health* **24**(1), 50–60 (2021). <https://doi.org/10.1016/j.jval.2020.10.004>
42. Rafferty, E.R., McDonald, W., Osgood, N.D., Qian, W., Doroshenko, A.: Seeking the optimal schedule for chickenpox vaccination in Canada: using an agent-based model to explore the impact of dose timing, coverage and waning of immunity on disease outcomes. *Vaccine* **38**(3), 521–529 (2020). <https://doi.org/10.1016/j.vaccine.2019.10.065>
43. Rahmandad, H., Sterman, J.: Heterogeneity and network structure in the dynamics of diffusion: comparing agent-based and differential equation models. *Manag. Sci.* **54**(5), 998–1014 (2008). <https://doi.org/10.1287/mnsc.1070.0787>
44. Railsback, S.F., Grimm, V.: Agent-Based and Individual-Based Modeling: A Practical Introduction. Princeton University Press, Princeton (2019)

45. Read, J.M., Keeling, M.J.: Disease evolution on networks: the role of contact structure. *Proc. R. Soc. Lond. B: Biol. Sci.* **270**(1516), 699–708 (2003). <https://doi.org/10.1098/rspb.2002.2305>
46. Richardson, G.P.: System Dynamics, pp. 807–810. Springer, New York (2001). https://doi.org/10.1007/1-4020-0611-X_1030
47. Richardson, G.P.: Core of System Dynamics. *System Dynamics: Theory and Applications* pp. 11–20 (2020)
48. Safarishahrbijari, A., Teyhouee, A., Waldner, C., Liu, J., Osgood, N.D.: Predictive accuracy of part. filt. in dyn. models supporting outbreak proj. *BMC Infect. Dis.* **17**(1), 1–12 (2017)
49. Smith, T., Rotondo, J., Desai, S., Deehan, H.: Pertussis: Pertussis surveillance in Canada: trends to 2012. *Can. Communicable Dis. Rep.* **40**(3), 21 (2014)
50. Statistics Canada: Data products, 2008 census (2008). <https://www150.statcan.gc.ca/n1/en/pub/82-224-x/2005000/5802980-eng.pdf?st=H2WZgClh>
51. Statistics Canada: Data products, 2016 census (2016). <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm>
52. Sterman, J.: Business Dynamics : Systems Thinking and Modeling for a Complex World. Title Match 2003: Murray Matheson. Irwin/McGraw-Hill, Boston (2000)
53. Tan, T., Dalby, T., Forsyth, K., Halperin, S.A., Heininger, U., Hozbor, D., Plotkin, S., Ulloa-Gutierrez, R., Von König, C.H.W.: Pertussis across the globe: recent epidemiologic trends from 2000 to 2013. *Pediatr. Infect. Dis. J.* **34**(9), e222–e232 (2015)
54. The AnyLogic Company: AnyLogic software. <https://www.anylogic.com>
55. Tian, Y., Osgood, N.: 15 things system dynamics can learn from software development. In: Proceedings of the 2012 International Conference of the System Dynamics Society, p. 18pp. System Dynamics Society, St. Gallen, Switzerland (2012)
56. Ulam, S.M.: Some ideas and prospects in biomathematics. *Ann. Rev. Biophys. Bioeng.* **1**(1), 277–292 (1972)
57. Weinstein, M.C., Russell, L.B., Gold, M.R., Siegel, J.E., et al.: Cost-Effectiveness in Health and Medicine. Oxford University Press, Oxford (1996)
58. Wiegers, K.E.: Peer Reviews in Software: A Practical Guide. Addison-Wesley Boston (2002)
59. World Health Organization: Pertussis (2019). <https://www.who.int/immunization/diseases/pertussis/en/>

Chapter 10

Mathematical Assessment of the Role of Interventions Against SARS-CoV-2



Salman Safdar and Abba B. Gumel

10.1 Introduction

COVID-19, the pneumonia-like illness that emerged out of Wuhan city in China late in December of 2019, has caused a devastating pandemic on a scale never before seen since the 1918/1919 influenza pandemic [48]. It has, as of April 2, 2023, caused over 683 million confirmed cases and over 6.8 million deaths globally [86]. The United States suffers the highest burden of the pandemic globally (with over 106 million confirmed cases and over 1.1 million deaths, as of April 2, 2023) [86]. The index case for COVID-19 was reported in the United States on January 21, 2020 [35]. For most parts of the year 2020, the control and mitigation efforts against SARS-CoV-2 in the United States were restricted to the use of non-pharmaceutical interventions, such as social distancing, quarantine of suspected cases, isolation of those with symptoms of SARS-CoV-2, the use of face coverings (i.e., face masks), community lockdowns, contact tracing, etc. [18, 57–60], until the Food and Drug Administration (FDA) provided Emergency Use Authorization (EUA) to two safe and highly efficacious vaccines (developed by Pfizer Inc. and Moderna Inc., respectively) in December of 2020 [25, 64]. Each of these two FDA-EUA vaccines was primarily administered in a two-dose regimen, within 3–4 weeks apart, and each offered an estimated protective efficacy against symptomatic COVID-19 infection of about 95% [50, 70]. Another vaccine, developed by Johnson & Johnson (administered as a single dose), received FDA-EUA in late February 2021 [81] (this vaccine has an estimated 75% efficacy in preventing severe/critical illness caused by

S. Safdar

School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ, USA
e-mail: ssafdar1@asu.edu

A. B. Gumel (✉)

Department of Mathematics, University of Maryland, College Park, MD, USA
e-mail: agumel@umd.edu

COVID-19 [68]). The rapid development and deployment of effective anti-COVID vaccines has played a vital role in minimizing and mitigating the burden of the pandemic, in jurisdictions with high coverage of these vaccines, around the world [4, 65].

This chapter focuses on using mathematical modeling approaches, coupled with rigorous qualitative data and data analytics together with computation, to assess the population-level impacts of the aforementioned three FDA-EUA vaccines on curtailing and mitigating the burden of the SARS-CoV-2 pandemic in the United States. Numerous clinical studies have shown that the efficacy of all the FDA-approved SARS-CoV-2 vaccines wane over time (with estimated waning time of about 9 months) [13, 14, 31, 61, 66]. Consequently, this chapter will specifically assess the population-level impact of the vaccination program (based on the three FDA-EUA vaccines), keeping in mind the waning efficacies of the approved vaccines, on the dynamics of the current predominant SARS-CoV-2 variant (Omicron) in the United States. The population-level impact of the use of face masks, as a singular intervention and its combination with vaccination, will also be assessed.

Numerous mathematical models, of various types (such as compartmental [18, 30, 40, 57–59, 69], agents-based [12, 23, 85], network [24, 77, 87], and statistical models [42, 73, 75]), have been formulated and used to gain insight and understanding into the transmission dynamics and control of the COVID-19 pandemic (with majority of these models being of the form of compartmental deterministic systems of nonlinear differential equations [18, 30, 40, 58, 59, 61]). In this chapter, a deterministic model will also be developed and used to study the dynamics of the COVID-19 pandemic. The model (which is relatively basic) will be rigorously analyzed to, among others, derive the threshold vaccination coverage needed to achieve herd immunity in the United States (the determination of this threshold provides the sufficient condition, in parameter space, for effectively curtailing and eliminating the pandemic in a population). Since the United States has been experiencing the brunt of the burden of the COVID-19 pandemic globally (i.e., the United States has so far recorded the highest number of confirmed cases and mortality) [16, 86], the emphasis of this chapter is on studying the dynamics of the COVID-19 pandemic (i.e., the current predominant Omicron variant) in the United States. Consequently, the model to be developed in this chapter will be parameterized using the daily new case data for the COVID-19 pandemic during the onset of the Omicron variant in the United States. The model will also be used to assess the population-level impact of the use of face masks (of various types and efficacy) on the disease dynamics. The chapter is organized as follows. The model for COVID-19 pandemic, in the presence of an imperfect vaccine, is formulated in Sect. 10.2. In addition to fitting the model using the daily new case data, the basic qualitative properties of the model are also presented in this section. The model is rigorously analyzed, with respect to the existence and asymptotic stability properties of its equilibria (disease-free and endemic), in Sect. 10.3. Rigorous analysis for, and numerical illustration of, the existence of the dynamic phenomenon of backward bifurcation in the model is also provided. Conditions for achieving community-wide vaccine-derived *herd immunity* are derived and global parameter sensitivity analyses

are also carried out in this section. Numerical simulations are reported in Sect. 10.4. The results of this chapter are discussed and summarized in Sect. 10.5.

10.2 Formulation of Vaccination Model for COVID-19

To develop the vaccination model for the transmission dynamics of COVID-19 in a population, the total population at time t , denoted by $N(t)$, is sub-divided into the mutually exclusive compartments of susceptible individuals ($S(t)$), fully vaccinated individuals ($V(t)$), exposed or latent individuals (i.e., newly infected individuals who are not yet infectious; $E(t)$), pre-symptomatic infectious individuals ($I_p(t)$), symptomatically infectious individuals ($I_s(t)$), asymptotically infectious individuals ($I_a(t)$), hospitalized individuals ($I_h(t)$), recovered individuals with natural immunity ($R_n(t)$), and recovered individuals with natural plus vaccine-derived immunity ($R_{nv}(t)$), so that:

$$N(t) = S(t) + V(t) + E(t) + I_p(t) + I_s(t) + I_a(t) + I_h(t) + R_n(t) + R_{nv}(t).$$

The model for COVID-19 dynamics in a population, which incorporates the use of any of the aforementioned imperfect vaccines, is given by the following deterministic system of nonlinear differential equations (where a dot represents differentiation with respect to time t) [66]. The flow diagram of the vaccination model is depicted in Fig. 10.1, and the description of the state variables and parameters of the vaccination model is tabulated in Tables 10.1 and 10.2, respectively:

$$\begin{cases} \dot{S} &= \Pi + \omega_v V - (\lambda + \xi_v + \mu)S, \\ \dot{V} &= \xi_v S - [(1 - \varepsilon_v)\lambda + \omega_v + \mu]V, \\ \dot{E} &= \lambda[S + (1 - \varepsilon_v)V + (1 - \varepsilon_n)R_n + (1 - \varepsilon_{nv})R_{nv}] - (\sigma_E + \mu)E, \\ \dot{I}_p &= \sigma_E E - (\sigma_p + \delta_p + \mu)I_p, \\ \dot{I}_s &= r\sigma_p I_p - (\phi_s + \gamma_s + \delta_s + \mu)I_s, \\ \dot{I}_a &= (1 - r)\sigma_p I_p - (\gamma_a + \delta_a + \mu)I_a, \\ \dot{I}_h &= \phi_s I_s - (\gamma_h + \delta_h + \mu)I_h, \\ \dot{R}_n &= \gamma_s I_s + \gamma_a I_a + \gamma_h I_h - [(1 - \varepsilon_n)\lambda + \xi_v + \mu]R_n, \\ \dot{R}_{nv} &= \xi_v R_n - [(1 - \varepsilon_{nv})\lambda + \mu]R_{nv}, \end{cases} \quad (10.1)$$

where the infection rate (or *force of infection*), λ , is given by:

$$\lambda = \frac{(\beta_p I_p + \beta_s I_s + \beta_a I_a + \beta_h I_h)}{N}, \quad (10.2)$$

with β_p , β_s , β_a , and β_h represent, respectively, the effective contact rates for pre-symptomatic (I_p), symptomatic (I_s), asymptomatic (I_a), and hospitalized (I_h) infectious individuals. In the vaccination model (10.1), Π is the rate of recruitment

Fig. 10.1 Flow diagram of the vaccination model (10.1)

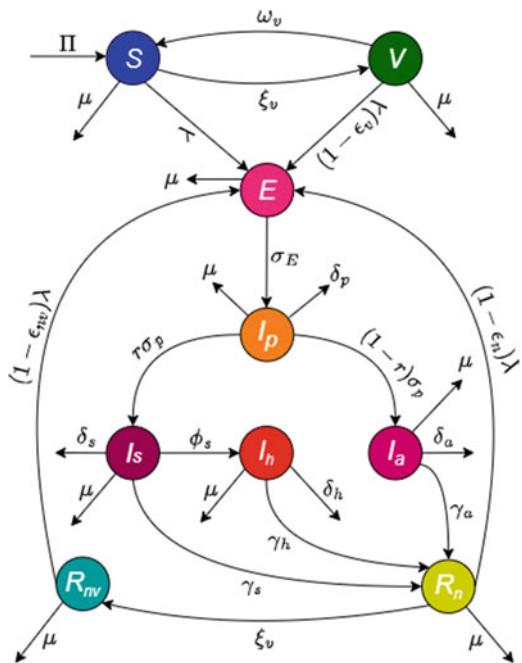


Table 10.1 Description of the state variables of the vaccination model (10.1) [66]

State variables	Description
S	Population of unvaccinated (wholly) susceptible individuals
V	Population of fully vaccinated susceptible individuals
E	Population of exposed (newly infected) individuals
I_p	Population of pre-symptomatic infectious individuals
I_s	Population of infectious individuals with clinical symptoms of the disease
I_a	Population of asymptomatically infectious individuals
I_h	Population of hospitalized individuals
R_n	Population of recovered individuals with natural immunity (i.e., unvaccinated)
R_{nv}	Population of recovered individuals with both natural and vaccine-derived immunity

of individuals into the population, ω_v is the vaccine waning rate for fully vaccinated individuals (i.e., the rate at which individuals in the V class revert to the wholly susceptible class, S), λ is the infection rate (defined in Eq. (10.2)), ξ_v is the *per capita* vaccination rate, and μ is the natural death rate. The parameter $0 < \varepsilon_v < 1$ is the average protective efficacy of the vaccine for fully vaccinated susceptible individuals (i.e., vaccine efficacy for individuals in the V class), while $0 \leq \varepsilon_n < 1$ is the efficacy of natural immunity to prevent recovered individuals (in the R_n class) from acquiring future SARS-CoV-2 infection and $0 < \varepsilon_{nv} \leq 1$ is the efficacy

Table 10.2 Description of the parameters of the vaccination model (10.1) [66]

Parameter	Description
Π	Recruitment rate
$\beta_p(\beta_s)(\beta_a)(\beta_h)$	Effective contact rate for individuals in the $I_p(I_s)(I_a)(I_h)$ compartment
ξ_v	<i>Per capita</i> vaccination rate
μ	Natural death rate
r	Proportion of pre-symptomatic individuals who show clinical symptoms of the disease
ω_v	Waning rate of fully vaccinated individuals
ε_v	Vaccine efficacy for fully vaccinated individuals
ε_n	Efficacy of natural immunity to prevent reinfection of recovered individuals in the R_n class
ε_{nv}	Efficacy of natural and vaccine-derived immunity to prevent reinfection of recovered individuals in the R_{nv} class
σ_E	Progression rate from exposed class to pre-symptomatic class
σ_p	Progression rate from pre-symptomatic class to either symptomatic or asymptomatic class
$\gamma_k(k = \{s, a, h\})$	Recovery rate for individuals in the I_s , I_a , and I_h class, respectively
ϕ_s	Hospitalization rate of individuals with clinical symptoms of the disease
$\delta_j(j = \{p, s, a, h\})$	Disease-induced mortality rate for individuals in the I_p , I_s , I_a , and I_h class, respectively

of natural and vaccine-derived immunity to prevent future SARS-CoV-2 infection of recovered individuals (in the R_{nv} class). Exposed individuals progress to the pre-symptomatic stage at the rate σ_E , and pre-symptomatic individuals progress to either become symptomatically infectious, at a rate $r\sigma_p$ (where $0 \leq r \leq 1$ is the proportion of these individuals that show clinical symptoms), or become asymptotically infectious, at the rate $(1 - r)\sigma_p$. Symptomatic individuals are hospitalized at a rate ϕ_s , and infectious individuals in stage I_k recover at a rate γ_k (with $k = \{s, a, h\}$). Finally, disease-induced mortality occurs in the I_j class at a rate δ_j ($j = \{p, s, a, h\}$) [66].

Some of the main assumptions made in formulating the vaccination model (10.1) are:

- (a) A homogeneously mixed population: it is assumed that the population is well-mixed, so that every member of the community is equally likely to mix with (and acquire infection from or transmit infection to) every other member of the community.
- (b) Vaccinated susceptible individuals (in the V class) are assumed to have received the full required doses (i.e., two doses for the Pfizer or the Moderna vaccine, one dose for the Johnson & Johnson vaccine), and that enough time has elapsed for the body to develop the full vaccine-derived immunity.
- (c) The three SARS-CoV-2 vaccines that received FDA Emergency Use Authorization (i.e., the Pfizer, Moderna, and Johnson & Johnson vaccines) are imperfect [26, 79, 81] (i.e., the vaccines offer partial protective immunity with efficacy

- $0 < \varepsilon_v < 1$), which wanes over time (at a rate ω_v) [13, 31]). In other words, vaccinated individuals can experience breakthrough infection [62, 80].
- (d) It is assumed that vaccine-derived immunity may wane over time in vaccinated individuals (V), resulting, ultimately, in reverting to the wholly susceptible class (S) [31].
 - (e) It is assumed, for mathematical tractability, that recovered individuals in the R_n and R_{nv} classes do not lose their natural (infection-acquired) immunity.
 - (f) Vaccination is only offered to wholly susceptible individuals or those who recovered naturally but their natural immunity has waned completely or recovered individuals who had acquired both natural and vaccine-derived immunity but their immunity has completely waned over time (i.e., the vaccines are not administered to individuals who are currently infected with SARS-CoV-2).
 - (g) It is assumed that pre-symptomatically infectious individuals do not recover while at this class (owing to the short average duration in this compartment). They recover only after transitioning to the symptomatic infectious class (at the rate $r\sigma_p$) or to the asymptomatically infectious class (at the rate $(1 - r)\sigma_p$). It is also assumed that individuals in the pre-symptomatic and asymptomatic infectious classes do not die from SARS-CoV-2 infection (so that δ_p and δ_a are set to zero in the numerical simulations) [53, 74, 83].

The vaccination model (10.1) extends numerous other (relatively basic) models for COVID-19 dynamics that incorporate the use of a vaccine, such as those in [30, 40], by *inter alia*:

- (a) Adding an epidemiological class for pre-symptomatic infectious individuals (the vaccination model in [30] does not explicitly account for disease transmission by pre-symptomatic infectious individuals).
- (b) Incorporating two classes for recovered individuals based on immunity status (i.e., recovered individuals with either natural or vaccine-derived immunity). Only one recovered compartment is considered in [30, 40].
- (c) Allowing for the reinfection of recovered individuals. This is not considered in [30, 40]. Furthermore, this study will contribute to the literature on the rigorous analyses of relatively basic vaccination models for COVID-19 by giving rigorous results for the existence and asymptotic stability of an endemic equilibrium of the (special case of) vaccination model (10.1).

10.2.1 Data Fitting and Parameter Estimation

The vaccination model (10.1) contains 22 parameters. Although the values of some of these parameters are known from the literature (as tabulated in Table 10.3), the values of some other parameters are unknown. Specifically, the values of the effective contact rate parameters (β_p , β_a , β_s , and β_h) are unknown. We fit the model with available data for COVID-19 for the United States and use the fitted

Table 10.3 Baseline values of the fixed parameters of the vaccination model (10.1) [66]

Parameter	Baseline value	Source
σ_E	$1/5 \text{ day}^{-1}$	[61]
σ_p	$1/2 \text{ day}^{-1}$	[49]
r	0.095 (dimensionless)	[84]
$(1 - r)$	0.905 (dimensionless)	[84]
γ_s	$1/10 \text{ day}^{-1}$	[31, 51]
γ_a	$1/5 \text{ day}^{-1}$	[44]
γ_h	$1/8 \text{ day}^{-1}$	[44]
ω_v	$1/274 \text{ day}^{-1}$	[13]
ϕ_s	$1/5 \text{ day}^{-1}$	[49]
ξ_v	$1.9 \times 10^{-5} \text{ day}^{-1}$	[61]
Π	$11,400 \text{ day}^{-1}$	[60]
μ	$3.4 \times 10^{-5} \text{ day}^{-1}$	[60]
δ_s	$4.9804 \times 10^{-5} \text{ day}^{-1}$	[66]
δ_h	$5.0 \times 10^{-5} \text{ day}^{-1}$	[60]
δ_p	0 day^{-1}	[53, 74]
δ_a	0 day^{-1}	[53, 83]
ε_v	0.85 (dimensionless)	[66]
ε_n	0.85 (dimensionless)	[66]
ε_{nv}	0.95 (dimensionless)	[66]

model to estimate the best values for the four unknown parameters. For fitting purposes, we use the daily new case data for the United States, obtained from the Johns Hopkins University COVID-19 repository [16], for the period between November 28, 2021 (when the Omicron variant first emerged in the United States), and March 23, 2022. The data fitting is done by splitting the data into two segments. The first segment of the data, from November 28, 2021, to February 23, 2022 (i.e., the region to the left of the dashed vertical cyan line in Fig. 10.2), was used to fit the model (10.1) and to estimate the unknown parameters [61, 66]. The second segment (from February 24, 2022, to March 23, 2022) was used to cross-validate the model [61, 66]. The model fitting was done using a standard nonlinear least squares approach, which involves using the inbuilt MATLAB’s minimization function (i.e., *lsqcurvefit*) to minimize the sum of the squared differences between each of the observed daily new case data points and the corresponding daily new case points obtained from the vaccination model (10.1) (i.e., $r\sigma_p I_p$).

Bootstrapping technique was used for the parameter estimation with 95% confidence intervals [3, 11, 58]. The process of bootstrapping involves producing a large collection of simulated data sets from a given data set by sampling from this given data set with replacement and then using each generated data set for the parameter estimation [60, 61]. The inbuilt bootstrapping function in MATLAB R2019b (i.e., *bootstrp*) was used to generate 10,000 bootstrap replicates to sample the residuals from the initial parameter estimation. Furthermore, the bootstrap data sets are generated by adding the re-sampled residuals to the best fit curve [61]. Finally, the vaccination model (10.1) was then fitted to each bootstrap data set to

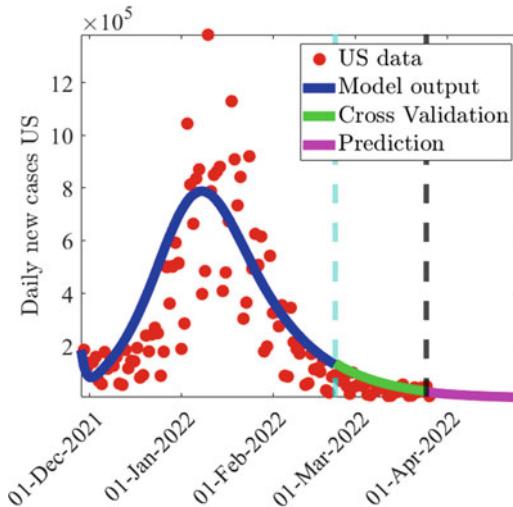


Fig. 10.2 Time series illustration of the least squares fit of the vaccination model (10.1), showing the model's output for the daily new cases in the United States (blue curve) compared to the observed confirmed daily new cases for the United States (red dots) from November 28, 2021, to February 23, 2022 (segment to the left of the dashed vertical cyan line), using the fixed and estimated (fitted) baseline parameter values given in Tables 10.3 and 10.4, respectively. The segment from February 24, 2022, to April 30, 2022 (i.e., solid green and magenta curves or the entire segment to the right of the dashed cyan vertical line), illustrates the performance of the vaccination model (10.1) in predicting the daily new COVID-19 cases in the United States [66]

create the bootstrap distribution of the estimated parameters. The inbuilt MATLAB R2019b function (i.e., *prctile*) is used for the bootstrap distribution to estimate the 95% credible intervals of the estimated parameters. Furthermore, it is significant to mention that fitting the model (10.1) to the daily new confirmed case data is much accurate data fitting approach which avoid mistakes that arise when cumulative case data is used for fitting the deterministic models [43, 61]. The values of the unknown parameters, which are estimated from the model fitting, are presented in Table 10.4.

The results obtained from fitting the vaccination model (10.1) with the observed daily new case data, depicted in Fig. 10.2, show a very good fit for the model output (blue curve) and the observed daily new case data (red dots) [66]. This figure also shows a very good fit for the cross-validation component of the fitting (green curve) of Fig. 10.2. This segment of Fig. 10.2 shows that the vaccination model (10.1) cross-validates the observed daily new case data for the period from February 23, 2022, to March 23, 2022, perfectly (solid green curve) [66]. The cross-validated model was used to make prediction for the trajectory of the pandemic for a 5-week period after the cross-validation period (March 24, 2022; as highlighted by the region to the right of the dashed vertical black line), as illustrated by the solid magenta curve in Fig. 10.2 [66].

Table 10.4 Baseline values of the four fitted (estimated) parameters (and their confidence intervals (CIs)) of the vaccination model (10.1), obtained by fitting the model with the observed daily new case COVID-19 data for the United States for the period November 28, 2021, to February 23, 2022 [66]

Parameter	Estimated value	95% confidence interval
β_p	0.2309078861597 day $^{-1}$	[0.2177887414779–0.2474537993681]
β_s	$9.9986277195284 \times 10^{-4}$ day $^{-1}$	[0.000100000000–0.0009999999991]
β_a	0.5429699023763 day $^{-1}$	[0.5342256853847–0.5523086685166]
β_h	$4.9989752860390 \times 10^{-5}$ day $^{-1}$	[0.000001000000–0.0000499999999]

10.2.2 Basic Qualitative Properties

Before carrying out the asymptotic analysis and numerical simulations of the vaccination model (10.1), it is instructive to explore its basic qualitative features with respect to its well-posedness (i.e., with respect to the non-negativity, boundedness, and invariance of its solutions). First of all, since the vaccination model (10.1) monitors the temporal dynamics of human populations, all its parameters are non-negative. It is convenient to define the following biologically feasible region for the vaccination model (10.1):

$$\Omega = \left\{ (S, V, E, I_p, I_s, I_a, I_h, R_n, R_{nv}) \in \mathbb{R}_+^9 : N(t) \leq \frac{\Pi}{\mu} \right\},$$

where $N(t)$ is the total population. It should be stated that since the vaccination model (10.1) monitors human populations, all its initial conditions are non-negative (i.e., $S(0) > 0$, $V(0) \geq 0$, $E(0) \geq 0$, $I_p(0) \geq 0$, $I_s(0) \geq 0$, $I_a(0) \geq 0$, $I_h(0) \geq 0$, $R_n(0) \geq 0$, $R_{nv}(0) \geq 0$). Furthermore, for the model to be mathematically and biologically meaningful, it is necessary that all solutions of the model remain non-negative for all non-negative initial conditions. That is, initial solutions of the model that start in the region Ω remain in Ω for all time $t > 0$ (i.e., Ω is positively invariant with respect to the vaccination model (10.1)). This result is rigorously established below.

Theorem 1 Consider the vaccination model (10.1) with non-negative initial conditions. The region Ω is positively invariant and attracts all solutions of the model (10.1).

Proof Adding all the equations of the model (10.1) gives [66]:

$$\dot{N} = \Pi - \mu N - \delta_p I_p - \delta_s I_s - \delta_a I_a - \delta_h I_h. \quad (10.3)$$

Since all the parameters of the model (10.1) are non-negative, it follows from Eq. (10.3) that:

$$\dot{N} \leq \Pi - \mu N. \quad (10.4)$$

Hence, if $N(t) > \frac{\Pi}{\mu}$, then $\dot{N} < 0$. Furthermore, by applying a standard comparison theorem [31, 47, 66] on (10.4), the following inequality holds:

$$N(t) \leq N(0)e^{-\mu t} + \frac{\Pi}{\mu} (1 - e^{-\mu t}).$$

Hence, if $N(0) \leq \frac{\Pi}{\mu}$, then $N(t) \leq \frac{\Pi}{\mu}$. If $N(0) > \frac{\Pi}{\mu}$ (which means that $N(0)$ is outside Ω) then $N(t) > \frac{\Pi}{\mu}$, for all $t > 0$ but with $\lim_{t \rightarrow \infty} N(t)$ (and this type of solution trajectory strives to enter the region Ω) [31]. Therefore, every solution of the vaccination model (10.1) with initial conditions in Ω remains in Ω for all time t . In other words, the region Ω is positively invariant and attracts all initial solutions of the vaccination model (10.1) [66]. \square

The epidemiological consequence of Theorem 1 is that it is sufficient to consider the dynamics of the flow generated by the vaccination model (10.1) in the invariant and bounded region Ω (since the model (10.1) is well-posed epidemiologically and mathematically in the feasible region Ω [36]). The existence and asymptotic stability properties of the equilibria of the model (10.1) will now be explored.

10.3 Existence and Asymptotic Stability of Equilibria

In this section, the vaccination model (10.1) will be rigorously analyzed to explore the conditions for the existence and asymptotic stability of its equilibria.

10.3.1 Disease-Free Equilibrium

The vaccination model (10.1) has a unique disease-free equilibrium (DFE) given by:

$$\begin{aligned} \mathcal{E}_0 &= \left(S^*, V^*, E^*, I_p^*, I_s^*, I_a^*, I_h^*, R_n^*, R_{nv}^* \right) \\ &= \left(\frac{\Pi(\mu + \omega_v)}{\mu(\mu + \xi_v + \omega_v)}, \frac{\Pi\xi_v}{\mu(\mu + \xi_v + \omega_v)}, 0, 0, 0, 0, 0, 0, 0 \right). \end{aligned}$$

10.3.1.1 Local Asymptotic Stability of DFE

The asymptotic stability property of the DFE (\mathcal{E}_0) will be explored using the *next-generation operator method* [15, 82]. Specifically, using the notation in [82], it follows that the associated non-negative matrix of new infection terms (F) and the M-matrix of the linear transition terms (V) are given, respectively, by [66]:

$$F = \begin{bmatrix} 0 & f_{11} & f_{12} & f_{13} & f_{14} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} K_1 & 0 & 0 & 0 & 0 \\ -\sigma_E & K_2 & 0 & 0 & 0 \\ 0 & -r\sigma_p & K_3 & 0 & 0 \\ 0 & -(1-r)\sigma_p & 0 & K_4 & 0 \\ 0 & 0 & -\phi_s & 0 & K_5 \end{bmatrix}, \quad (10.5)$$

where

$$\begin{aligned} f_{11} &= \beta_p \left[\frac{S^* + (1 - \varepsilon_v)V^*}{N^*} \right], \quad f_{12} = \beta_s \left[\frac{S^* + (1 - \varepsilon_v)V^*}{N^*} \right], \\ f_{13} &= \beta_a \left[\frac{S^* + (1 - \varepsilon_v)V^*}{N^*} \right], \quad f_{14} = \beta_h \left[\frac{S^* + (1 - \varepsilon_v)V^*}{N^*} \right], \quad K_1 = \sigma_E + \mu, \\ K_2 &= \sigma_p + \mu + \delta_p, \quad K_3 = \phi_s + \gamma_s + \mu + \delta_s, \quad K_4 = \gamma_a + \mu + \delta_a, \quad \text{and} \quad K_5 = \gamma_h + \mu + \delta_h. \end{aligned}$$

Let f_v represent the proportion of wholly susceptible individuals that are fully vaccinated at disease-free equilibrium. In other words,

$$f_v = \frac{V^*}{N^*} = \frac{\xi_v}{\mu + \xi_v + \omega_v}. \quad (10.6)$$

It follows, based on the computation of the next-generation matrices above [15, 82], that the *vaccination reproduction number* (or the *control reproduction number*) of the vaccination model (10.1), denoted by \mathbb{R}_{cv} , is given by (where ρ is the spectral radius):

$$\mathbb{R}_{cv} = \rho(FV^{-1}) = \mathbb{R}_0 \left(1 - \varepsilon_v \frac{V^*}{N^*} \right) = \mathbb{R}_0 (1 - \varepsilon_v f_v), \quad (10.7)$$

where f_v is as defined in Eq. (10.6) and

$$\mathbb{R}_0 = \mathbb{R}_{cv}|_{\varepsilon_v=V=0}, \quad (10.8)$$

is the *basic reproduction number* of the vaccination model (10.1). It can be shown (by applying the next-generation operator method on the model (10.1) in the absence of vaccination) that:

$$\mathbb{R}_0 = \mathbb{R}_{0p} + \mathbb{R}_{0s} + \mathbb{R}_{0a} + \mathbb{R}_{0h}, \quad (10.9)$$

where

$$\begin{aligned}\mathbb{R}_{0p} &= \beta_p \left(\frac{S^*}{N^*} \right) \left(\frac{\sigma_E}{\sigma_E + \mu} \right) \left(\frac{1}{\sigma_p + \delta_p + \mu} \right), \\ \mathbb{R}_{0s} &= \beta_s \left(\frac{S^*}{N^*} \right) \left(\frac{\sigma_E}{\sigma_E + \mu} \right) \left(\frac{r\sigma_p}{\sigma_p + \delta_p + \mu} \right) \left(\frac{1}{\phi_s + \gamma_s + \delta_s + \mu} \right), \\ \mathbb{R}_{0a} &= \beta_a \left(\frac{S^*}{N^*} \right) \left(\frac{\sigma_E}{\sigma_E + \mu} \right) \left(\frac{(1-r)\sigma_p}{\sigma_p + \delta_p + \mu} \right) \left(\frac{1}{\gamma_a + \delta_a + \mu} \right),\end{aligned}$$

and

$$\begin{aligned}\mathbb{R}_{0h} &= \beta_h \left(\frac{S^*}{N^*} \right) \left(\frac{\sigma_E}{\sigma_E + \mu} \right) \left(\frac{r\sigma_p}{\sigma_p + \delta_p + \mu} \right) \left(\frac{\phi_s}{\phi_s + \gamma_s + \delta_s + \mu} \right) \\ &\quad \times \left(\frac{1}{\gamma_h + \delta_h + \mu} \right),\end{aligned}$$

are the constituent reproduction numbers for the transmission of the disease by infectious individuals in the pre-symptomatic, symptomatic, asymptomatic, and hospitalized classes, respectively. It is worth stating that, while the basic reproduction number (\mathbb{R}_0) measures the average number of new SARS-CoV-2 cases generated by a typical infectious individual if introduced in a completely susceptible population, the control reproduction number (\mathbb{R}_{cv}) measures the average number of new SARS-CoV-2 cases generated by a typical infectious individual introduced into a population where a certain proportion of the wholly susceptible population is fully vaccinated (with any of the three aforementioned FDA-approved vaccines).

The asymptotic stability result below follows from Theorem 2 of [82]:

Theorem 2 *The disease-free equilibrium (\mathcal{E}_0) of the vaccination model (10.1) is locally asymptotically stable (LAS) if $\mathbb{R}_{cv} < 1$, and unstable if $\mathbb{R}_{cv} > 1$.*

The epidemiological implication of Theorem 2 is that a small influx of SARS-CoV-2 cases will not generate a large outbreak in the community if the control reproduction number (\mathbb{R}_{cv}) is brought to, and maintained at, a value less than unity [66]. In other words, the vaccination program implemented in the United States can lead to the effective control of the SARS-CoV-2 pandemic if it can result in reducing (and maintaining) the control reproduction number to a value less than one, provided the initial number of infectious individuals introduced into the population is small enough.

10.3.1.2 Existence of Backward Bifurcation

Certain epidemiological mechanisms associated with the spread and control of infectious diseases are known to induce *backward bifurcation*, a dynamic phenomenon characterized by the coexistence of two stable attractors (namely, the

stable disease-free equilibrium and a stable endemic equilibrium) when the associated reproduction number of the model is less than one [5, 27–29, 39, 41]. The epidemiological implication of the presence of a backward bifurcation in the transmission dynamics of an infectious disease is that bringing (and maintaining) the reproduction number of the model to a value less than one, while necessary, may not be sufficient to lead to the elimination of the disease [29, 39, 41]. One common cause of backward bifurcation in disease transmission models is the use of an imperfect vaccine [19, 20, 29]. Since the vaccination model (10.1) uses an imperfect vaccine, it is instructive to explore the likelihood (or derive the conditions for the occurrence) of a backward bifurcation in its transmission dynamics. This is done below. In particular, we claim the following result:

Theorem 3 *The vaccination model (10.1) undergoes a backward bifurcation at $\mathbb{R}_{cv} = 1$ whenever the associated bifurcation coefficients (denoted by a and b and given by Eqs. (10.26) and (10.27) in Appendix 1) are positive (or, equivalently, when Inequality (10.30) holds).*

The proof of Theorem 3, based on using the center manifold theory [8, 9, 17, 33, 82], is given in Appendix 1. Figure 10.3 depicts the associated backward bifurcation diagram for the vaccination model (10.1). It should be mentioned that, for computational convenience (in generating the bifurcation diagrams), we set, without loss of generality, the eigenvectors v_1 and v_3 in (10.22) (given in Appendix 1) to one. Similarly, we set the eigenvectors w_1 , w_7 , and w_9 in (10.23) to unity.

The epidemiological implication of Theorem 3 is that the ability of the intervention and mitigation programs implemented (vaccination and face mask usage in this case) to bring (and maintain) the control reproduction number, \mathbb{R}_{cv} , to a value less than one, while necessary, is no longer sufficient for the elimination of the SARS-CoV-2 pandemic in the community. Such control (within the bistability region in Fig. 10.3) is now dependent on the size of the initial sub-populations of the model. Specifically, initial conditions of the model (10.1) that lie below the stable manifold of the saddle point (the separatrix which separates the basin of attraction of the stable endemic equilibrium and that of the stable disease-free equilibrium) will converge to the disease-free equilibrium, while those that lie above the separatrix will converge to an endemic equilibrium point. It follows from Fig. 10.3 that, in order to be outside the backward bifurcation region, the intervention and mitigation measures implemented in the community would need to be ramped up to further reduce the control reproduction number below one (and outside the bistability region). Thus, the presence of a backward bifurcation in the transmission dynamics of a disease makes its effective control more difficult (since it imposes greater requirement in terms of the efficacy and coverage of interventions) [29, 33].

To ensure that the effective control of the disease (or its elimination) is independent of the initial sizes of the sub-populations of the model, it is necessary that the disease-free equilibrium is proved to be globally asymptotically stable when the reproduction number of the model is less than one. This is explored in Sect. 10.3.1.3, for two special cases of the vaccination model (10.1).

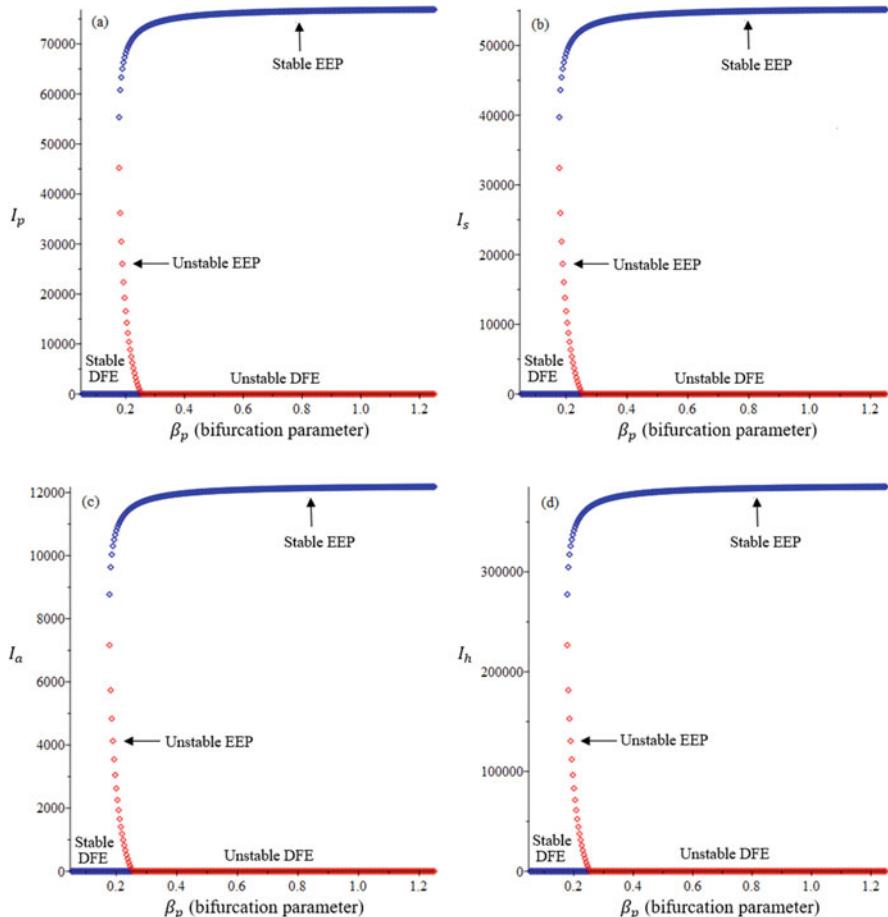


Fig. 10.3 Backward bifurcation diagram for the vaccination model (10.1), showing the profiles of the population of (a) pre-symptomatically infectious individuals (I_p), (b) symptomatic individuals (I_s), (c) asymptotically infectious individuals (I_a) and (d) hospitalized individuals (I_h), as a function of the bifurcation parameter β_p . Parameter values used are: $\Pi = 20,000$, $\omega_v = 0.00004$, $\mu = 0.001$, $\xi_v = 0.002$, $\varepsilon_v = 0.8$, $\varepsilon_n = 0.08$, $\varepsilon_{nv} = 0.08$, $\sigma_E = 0.2$, $\sigma_p = 0.98$, $\phi_s = 0.95$, $\gamma_s = 0.12$, $\gamma_a = 0.12$, $\gamma_h = 0.12$, $\delta_p = 0.0095$, $\delta_a = 0.0095$, $\delta_s = 0.015$, $\delta_h = 0.015$, $r = 0.9921$, $d_1 = 1.5$, $d_2 = 0.75$, $d_3 = 1$. With this arbitrary set of parameter values, the values of the associated backward bifurcation coefficients (denoted by a and b , and given in Appendix 1) are $a = 6.3833 \times 10^{-6} > 0$ and $b = 0.17481 > 0$, respectively. Furthermore, $\beta_p^* = 0.24010$ and $\mathbb{R}_{cv} = 1$. Apart from the efficacies (i.e., ε_v , ε_n , and ε_{nv}), scaling factors (i.e., d_1 , d_2 , and d_3), and the proportion “ r ”, which are dimensionless, all the other parameters have unit of per day

10.3.1.3 Global Asymptotic Stability of DFE: Special Cases

In this section, we explore extending the result in Theorem 2 to prove the global asymptotic stability of the DFE for two special cases of the vaccination model (10.1), as follows [66].

Special Case 1

Consider, first of all, the special case of the model (10.1) where the vaccines administered in the population are assumed to offer perfect protective efficacy against the original strain of the pandemic. That is, consider the model (10.1) with $\varepsilon_v = 1$. This assumption is plausible, for instance, in the case of the Pfizer or Moderna vaccine (with each being almost 95% effective against the original SARS-CoV-2 strain) [51, 63]. Furthermore, for mathematical tractability, it is assumed, for this special case, that natural immunity is perfect against reinfection (so that $\varepsilon_n = \varepsilon_{nv} = 1$).

For this special case of the vaccination model (10.1), it can be seen that the associated next-generation matrix of new infection terms, denoted by \tilde{F} , is given by (note that, for this special case, the next-generation matrix of linear transition terms, V , remains the same, as defined by Eq. (10.5). Furthermore, $N^* = \Pi/\mu$):

$$\tilde{F} = \begin{bmatrix} 0 \beta_p \left(\frac{S^*}{N^*} \right) \beta_s \left(\frac{S^*}{N^*} \right) \beta_a \left(\frac{S^*}{N^*} \right) \beta_h \left(\frac{S^*}{N^*} \right) \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \quad (10.10)$$

The control reproduction number of this special case of the basic model, denoted by $\tilde{\mathbb{R}}_{cv}$, is given by:

$$\tilde{\mathbb{R}}_{cv} = \rho(\tilde{F}V^{-1}) = \mathbb{R}_{cv}|_{\varepsilon_v=1}. \quad (10.11)$$

We claim the following result:

Theorem 4 Consider the special case of the vaccination model (10.1) with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ and $\tilde{\mathbb{R}}_{cv} \leq (1 - f_v) < 1$ (with f_v as defined in Eq. (10.6)). The disease-free equilibrium (\mathcal{E}_0) of the special case of the model is globally asymptotically stable in Ω whenever $\tilde{\mathbb{R}}_{cv} < 1$.

The proof of Theorem 4, based on using Lyapunov function [7, 30, 38], is given in Appendix 2. The result of Theorem 4 is numerically illustrated in Fig. 10.4, where all initial conditions of the special case of the model converged to the disease-free equilibrium when the associated control reproduction number, $\tilde{\mathbb{R}}_{cv}$, is less than one. The epidemiological implication of Theorem 4 is that, for the special case of the vaccination model (10.1) with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$, the COVID-19 pandemic can be eliminated in the United States if the threshold quantity, $\tilde{\mathbb{R}}_{cv}$, can be brought to (and maintained at) a value less than one. In other words, for the aforementioned special case of the model, having $\tilde{\mathbb{R}}_{cv} < 1$ is necessary and sufficient for the effective control (or elimination) of the pandemic in the United States. Hence, implementing a vaccination program that can bring (and maintain) $\tilde{\mathbb{R}}_{cv}$ to a value less than one will result in the elimination of the pandemic in the United States.

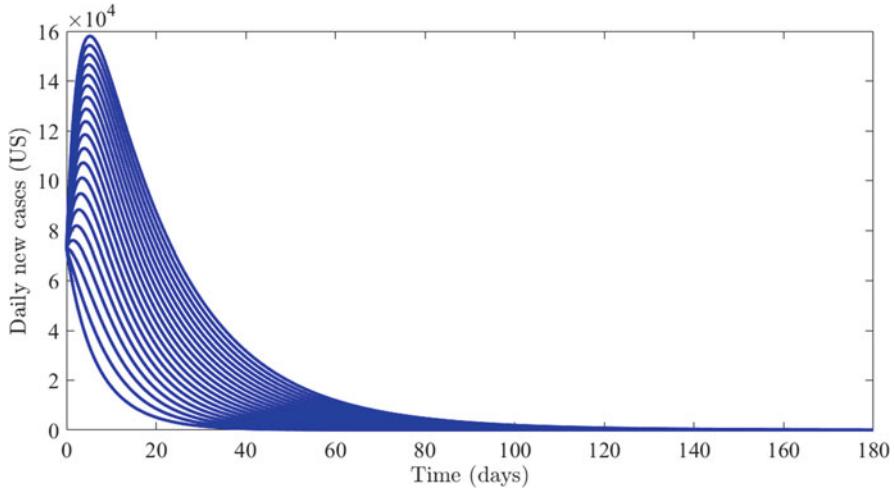


Fig. 10.4 Simulations of the special case of the vaccination model (10.1), for the average number of daily new cases in the United States as a function of time (in days), showing convergence of initial conditions to the disease-free equilibrium (DFE) when $\tilde{\mathbb{R}}_{cv} < 1$. The values of the parameters used in these simulations are as given by their baseline values given in Tables 10.3 and 10.4, with $\beta_a = 0.1542 \text{ day}^{-1}$. With this set of parameter values, $\tilde{\mathbb{R}}_{cv} = 0.3824 < 1$

Table 10.5 Effect of efficacies of vaccine-derived (ε_v), natural (ε_n), and combined natural and vaccine-derived (ε_{nv}) on the likelihood of the occurrence of backward bifurcation in the vaccination model (10.1), as measured by the values of the associated backward bifurcation coefficients, a and b (given in Appendix 1), and the values of β_p^* . Parameter values (chosen arbitrarily for illustrative purposes) used in generating this table are: $\Pi = 20,000$, $\omega_v = 0.00037$, $\mu = 0.000034$, $\xi_v = 0.0004277$, $\sigma_E = 0.2$, $\sigma_p = 0.5$, $\phi_s = 0.15$, $\gamma_s = 0.2$, $\gamma_a = 0.125$, $\gamma_h = 0.12$, $\delta_p = 0$, $\delta_a = 0$, $\delta_s = 0.0000498$, $\delta_h = 0.00005$, $r = 0.152$, $d_1 = 1.5$, $d_2 = 0.75$, $d_3 = 1$, $\beta_s = d_1 \times \beta_p$, $\beta_a = d_2 \times \beta_p$, $\beta_h = d_3 \times \beta_p$ and various values of ε_v , ε_n , and ε_{nv} . This set of parameter values is used to compute the corresponding values of the bifurcation parameter, β_p^* . Furthermore, for this set of parameter values, $\tilde{\mathbb{R}}_{cv} = 1$. Apart from the efficacies (i.e., ε_v , ε_n , and ε_{nv}), scaling factors (i.e., d_1 , d_2 , and d_3), and the proportion “ r ,” which are dimensionless, all the other parameters have unit of *per day*. Notation a^\dagger denotes the bifurcation coefficient, a , while b^\dagger denotes the bifurcation coefficient, b

Immunity efficacy	Value of a^\dagger	Value of b^\dagger	Value of β_p^*
$\varepsilon_v = 0.50, \varepsilon_n = 0.50, \varepsilon_{nv} = 0.50$	$9.72721 \times 10^{-7} > 0$	$0.0330300 > 0$	0.21395
$\varepsilon_v = 0.85, \varepsilon_n = 0.85, \varepsilon_{nv} = 0.95$	$3.41948 \times 10^{-7} > 0$	$0.0991100 > 0$	0.23200
$\varepsilon_v = 0.90, \varepsilon_n = 0.90, \varepsilon_{nv} = 0.95$	$2.33118 \times 10^{-7} > 0$	$0.0660700 > 0$	0.23450
$\varepsilon_v = 0.95, \varepsilon_n = 0.95, \varepsilon_{nv} = 0.95$	$1.19874 \times 10^{-7} > 0$	$0.0330700 > 0$	0.23750
$\varepsilon_v = 1.00, \varepsilon_n = 1.00, \varepsilon_{nv} = 1.00$	$-2.723 \times 10^{-14} < 0$	$3.1 \times 10^{-9} > 0$	0.24010

It is worth mentioning that substituting $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ into the expressions for the backward bifurcation coefficients (a and b) in Appendix 1, and simplifying, shows that $a = -2.7089 \times 10^{-14} < 0$ and $b = 3.1 \times 10^{-9} > 0$, as tabulated in Table 10.5. Thus, it follows from Item (i) of Theorem 4.1 in [9] that, unlike the

full model (10.1), the special case of the model with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ will not undergo backward bifurcation at $\tilde{\mathbb{R}}_{cv} = 1$ (this is in line with the global asymptotic stability result proved for the disease-free equilibrium of the special case of the model in Theorem 4). In fact, this special case of the model undergoes a forward bifurcation at $\tilde{\mathbb{R}}_{cv} = 1$ (as depicted in Fig. 10.5), and no endemic equilibrium exists when $\tilde{\mathbb{R}}_{cv} < 1$. Hence, this study shows identifies a sufficient condition for the emergence of the dynamic behavior of backward bifurcation in the transmission dynamics of SARS-CoV-2, namely, the imperfect nature of the vaccine-derived and natural immunity (to prevent the acquisition of SARS-CoV-2 infection by vaccinated susceptible individuals and the reinfection of recovered individuals).

Table 10.5 shows that when the baseline values of the efficacies of the vaccine-derived immunity ($\varepsilon_v = 0.85$), natural immunity ($\varepsilon_n = 0.85$), and natural and vaccine-derived immunity ($\varepsilon_{nv} = 0.95$), given in Table 10.3, are substituted into the expressions for the associated backward bifurcation coefficients (a and b in Appendix 1), the values of these coefficients become $a = 3.41948 \times 10^{-7} > 0$ and $b = 0.09911 > 0$, respectively (see Table 10.5). In other words, it follows from Item (i) of Theorem 4.1 in [9] that backward bifurcation will occur if the baseline values of the parameters of the model (10.1), tabulated in Table 10.3, are used. Thus, this study shows that backward bifurcation is, indeed, a realistic feature in the transmission dynamics of SARS-CoV-2 in a population that uses imperfect vaccines and where natural (and combined natural and vaccine-derived) immunity does not offer perfect protection against reinfection. Table 10.5 further shows that the likelihood of a backward bifurcation occurring increases as the values of the parameters related to the vaccine-derived (ε_v) and natural immunity (ε_n and ε_{nv}) increase towards one (note that the likelihood of backward bifurcation decreases with decreasing values of the bifurcation coefficients, a and b ; and backward bifurcation does not occur when the coefficient a further decreases to values less than zero, in line with Item (i) of Theorem 4.1 in [9]). In other words, this study shows that the phenomenon of backward bifurcation is more likely to occur in the community if vaccines with lower protective efficacy are used and if the efficacy of natural and combined natural and vaccine-derived immunity to prevent reinfection in the community is low.

Special Case 2

The global asymptotic stability of the disease-free equilibrium of the vaccination model (10.1) can also be established for another special case of the model in the absence of disease-induced mortality and recovered individuals do not acquire SARS-CoV-2 reinfection. That is, we consider the special case of the vaccination model (10.1) with $\delta_p = \delta_s = \delta_a = \delta_h = 0$, and $\varepsilon_n = \varepsilon_{nv} = 1$. The assumption for having negligible disease-induced mortality is reasonable owing to the fact that (a) this study focuses on the SARS-CoV-2 dynamics in the United States during the period when Omicron is the predominant variant (i.e., starting from November 28, 2021) and (b) Omicron is far less fatal than the SARS-CoV-2 variants that preceded it (particularly Delta) [61, 66].

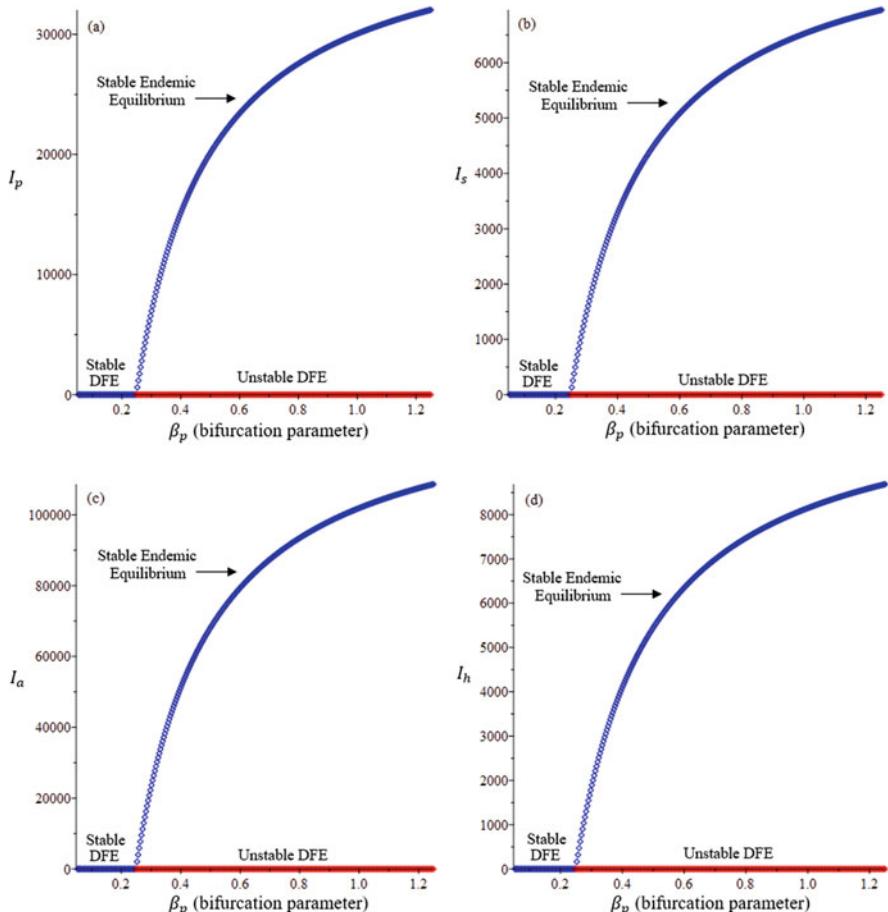


Fig. 10.5 Forward bifurcation diagram for the vaccination model (10.1), showing the profiles of the population of (a) pre-symptomatically infectious individuals (I_p), (b) symptomatic individuals (I_s), (c) asymptomatically infectious individuals (I_a), and (d) hospitalized individuals (I_h) as a function of the bifurcation parameter β_p . Parameter values used are: $\Pi = 20000$, $\omega_v = 0.00037$, $\beta_p = 0.24010$, $\beta_s = 0.36022$, $\beta_a = 0.18011$, $\beta_h = 0.24010$, $\mu = 0.000034$, $\xi_v = 0.0004277$, $\varepsilon_v = 1$, $\varepsilon_n = 1$, $\varepsilon_{nv} = 1$, $\sigma_E = 0.2$, $\sigma_p = 0.5$, $\phi_s = 0.15$, $\gamma_s = 0.2$, $\gamma_a = 0.125$, $\gamma_h = 0.12$, $\delta_p = 0$, $\delta_a = 0$, $\delta_s = 0.0000498$, $\delta_h = 0.00005$, $r = 0.152$, $d_1 = 1.5$, $d_2 = 0.75$, $d_3 = 1$. With this arbitrary set of parameter values, the values of the associated backward bifurcation coefficients (denoted by a and b and given in Appendix 1) are $a = -2.7231 \times 10^{-14} > 0$ and $b = 3.1 \times 10^{-9} > 0$, respectively. Furthermore, $\beta_p^* = 0.24010$ and $\mathbb{R}_{cv} = 1$. Apart from the efficacies (i.e., ε_v , ε_n , and ε_{nv}), scaling factors (i.e., d_1 , d_2 , and d_3), and the proportion “ r ”, which are dimensionless, all the other parameters have unit of per day

Setting $\delta_p = \delta_s = \delta_a = \delta_h = 0$ into the vaccination model (10.1), and adding all the resulting equations, shows that $\frac{dN}{dt} = \Pi - \mu N$, from which it follows that $N(t) \rightarrow \frac{\Pi}{\mu}$ as $t \rightarrow \infty$. From now on, the total population at time t , $N(t)$, will be replaced by its limiting value, $N^* = \Pi/\mu$ (i.e., the standard incidence formulation for the infection rate is now replaced by a mass action incidence). Consider the following feasible region for this (second) special case of the vaccination model (10.1):

$$\Omega_{**} = \{(S, V, E, I_p, I_s, I_a, I_h, R_n, R_{nv}) \in \Omega : S \leq S^*, V \leq V^*\}. \quad (10.12)$$

It can be shown that the region Ω_{**} is positively invariant and attracting with respect to this second special case of the vaccination model (10.1) (see 10.5 for the proof). Furthermore, it is convenient to define the following threshold quantity:

$$\hat{\mathbb{R}}_{cv} = \mathbb{R}_{cv}|_{\delta_p=\delta_s=\delta_a=\delta_h=0}. \quad (10.13)$$

We claim the following result:

Theorem 5 *Consider the special case of the vaccination model (10.1) in the absence of disease-induced mortality (i.e., $\delta_p = \delta_s = \delta_a = \delta_h = 0$) and no reinfection of recovered individuals (i.e., $\varepsilon_n = \varepsilon_{nv} = 1$). The disease-free equilibrium of this special case of the model (\mathcal{E}_0) is globally asymptotically stable in Ω_{**} whenever $\hat{\mathbb{R}}_{cv} < 1$.*

The proof of Theorem 5, based on using a comparison theorem, is given in Appendix 3.

10.3.2 Existence and Stability of Endemic Equilibria: Special Case

In this section, the possible existence and asymptotic stability of endemic (positive) equilibria (i.e., equilibria where at least one of the infected components is positive) of the vaccination model (10.1) will be explored for a special case. Specifically, we consider the special case where protective efficacy of the vaccines against primary infection and reinfection is 100% (i.e., $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$), no waning of vaccine-derived immunity (i.e., $\omega_v = 0$), and no disease-induced mortality (i.e., $\delta_p = \delta_s = \delta_a = \delta_h = 0$). For this special case of the vaccination model (10.1), the associated vaccination reproduction number is defined as follows:

$$\tilde{\mathbb{R}}_v = \mathbb{R}_{cv}|_{\delta_p=\delta_s=\delta_a=\delta_h=0, \varepsilon_v=1}. \quad (10.14)$$

10.3.2.1 Existence

Let $E_1 = (S^{**}, V^{**}, E^{**}, I_p^{**}, I_s^{**}, I_a^{**}, I_h^{**}, R_n^{**}, R_{nv}^{**})$ represent any arbitrary (positive) endemic equilibrium point (EEP) of the vaccination model (10.1), with $N^{**} = S^{**} + V^{**} + E^{**} + I_p^{**} + I_s^{**} + I_a^{**} + I_h^{**} + R_n^{**} + R_{nv}^{**}$. Consider the vaccination model (10.1) with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ and $\omega_v = \delta_p = \delta_s = \delta_a = \delta_h = 0$ and $\tilde{\mathbb{R}}_v > 1$. Solving the equations of this special case of the vaccination model at steady state, and simplifying, gives the following steady-state expressions:

$$\begin{aligned} S^{**} &= \frac{\Pi}{\lambda^{**} + c_1}, & V^{**} &= \frac{\xi_v S^{**}}{c_2}, & E^{**} &= \frac{\lambda^{**} S^{**}}{c_3}, & I_p^{**} &= \frac{\sigma_E \lambda^{**} S^{**}}{c_3 c_4}, \\ I_s^{**} &= \frac{r \sigma_p \sigma_E \lambda^{**} S^{**}}{c_3 c_4 c_5}, & I_a^{**} &= \frac{(1-r) \sigma_p \sigma_E \lambda^{**} S^{**}}{c_3 c_4 c_6}, \\ I_h^{**} &= \frac{\phi_s r \sigma_p \sigma_E \lambda^{**} S^{**}}{c_3 c_4 c_5 c_7}, & R_n^{**} &= \frac{\sigma_E \lambda^{**} S^{**} P_1}{c_3}, & R_{nv}^{**} &= \frac{\xi_v \sigma_E \lambda^{**} S^{**} P_1}{\mu c_3}, \end{aligned} \quad (10.15)$$

where

$$c_1 = \xi_v + \mu, c_2 = \omega_v + \mu, c_3 = \sigma_E + \mu, c_4 = \sigma_p + \mu,$$

$$c_5 = \phi_s + \gamma_s + \mu, c_6 = \gamma_a + \mu,$$

$$c_7 = \gamma_h + \mu, P_1 = \frac{r \sigma_p}{c_5} + \frac{(1-r) \sigma_p}{c_6} + \frac{\phi_s r \sigma_p}{c_5 c_7},$$

$$\tilde{P} = \frac{\xi_v}{c_2} + \frac{1}{c_3} + \frac{\sigma_E}{c_3 c_4} + \frac{r \sigma_E \sigma_p}{c_3 c_4 c_5} + \frac{(1-r) \sigma_E \sigma_p}{c_3 c_4 c_6} + \frac{r \phi_s \sigma_E \sigma_p}{c_3 c_4 c_5 c_6} + \frac{\sigma_E P_1}{c_3} + \frac{\xi_v \sigma_E P_1}{\mu c_3},$$

$$P_2 = 1 + P_1 + \tilde{P}$$

with

$$\lambda^{**} = \frac{(\beta_p I_p^{**} + \beta_s I_s^{**} + \beta_a I_a^{**} + \beta_h I_h^{**})}{N^{**}}. \quad (10.16)$$

It follows, by substituting the expressions for $I_p^{**}, I_s^{**}, I_a^{**}$, and I_h^{**} from Eq. (10.15) into Eq. (10.16) and extensive algebraic manipulations, that:

$$\lambda^{**} = \frac{(\tilde{\mathbb{R}}_v - 1)}{P_2}, \quad (10.17)$$

from which it follows that $\lambda^{**} > 0$ whenever $\tilde{\mathbb{R}}_v > 1$. Hence, the special case of the model has a unique endemic equilibrium point (EEP) whenever $\tilde{\mathbb{R}}_v > 1$. The components of this unique endemic equilibrium point can be obtained by substituting (10.17) into the expressions for $S^{**}, V^{**}, E^{**}, I_p^{**}, I_s^{**}, I_a^{**}, I_h^{**}, R_n^{**}$, and R_{nv}^{**} , given in Eq. (10.15). Thus, we have proved the following result:

Theorem 6 *The special case of the vaccination model (10.1), with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ and $\omega_v = \delta_p = \delta_s = \delta_a = \delta_h = 0$, has a unique endemic (positive) equilibrium, given by E_1 , whenever $\tilde{\mathbb{R}}_v > 1$.*

10.3.2.2 Local Asymptotic Stability

The local asymptotic property of the unique endemic equilibrium of the special case of the model (10.1) (which exists whenever $\tilde{\mathbb{R}}_v > 1$, as shown in Theorem 6) will now be explored. We claim the following result:

Theorem 7 *The unique endemic equilibrium point (\tilde{E}_1) of the special case of the vaccination model (10.1), with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ and $\omega_v = \delta_p = \delta_s = \delta_a = \delta_h = 0$ (which exists when $\tilde{\mathbb{R}}_v > 1$), is locally asymptotically stable whenever $\tilde{\mathbb{R}}_v > 1$.*

The proof of Theorem 7, based on using a Krasnoselskii sub-linearity argument [21, 37, 55, 67], is given in Appendix 4. The epidemiological implication of Theorem 7 is that, for the special case of the vaccination model (10.1) with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ and $\omega_v = \delta_p = \delta_s = \delta_a = \delta_h = 0$, the disease will persist in the population (when the associated reproduction $\tilde{\mathbb{R}}_v$ exceeds one) if the initial sizes of the sub-populations of the model are in the basin of attraction of the endemic equilibrium. The epidemiological implication of Theorems 6 and 7 is that, for the special case of the vaccination model (with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ and $\omega_v = \delta_p = \delta_s = \delta_a = \delta_h = 0$), the disease will persist, if the initial population sizes are in the basin of attraction of the unique endemic equilibrium, provided the value of the associated reproduction threshold ($\tilde{\mathbb{R}}_v$) exceeds unity. Figure 10.6 depicts a time series of initial solutions of the special case of the model for the case where $\tilde{\mathbb{R}}_v > 1$ showing convergence of all the initial solutions to the unique endemic equilibrium, in line with Theorem 7.

10.3.3 Vaccine-Induced Herd Immunity Threshold

Herd immunity, which is a measure of the minimum percentage of the number of susceptible individuals in a community that need to be protected against infection in order to eliminate community transmission of an infectious disease, can be attained through two main ways, namely, natural immunity route (following natural recovery from infection with the disease) or by vaccination (which is widely considered to be the safest and the fastest way) [1, 2]. For vaccine-preventable diseases, such as COVID-19, it is not practically possible to vaccinate every susceptible individual in the community due to various reasons, such as infants, individuals who are pregnant, breastfeeding women, individuals with certain underlying medical conditions, or those who are unwilling to be vaccinated for COVID-19 due to some other reasons [40, 66]. Thus, it is critical to know what minimum proportion of the susceptible population that need to be vaccinated in order to protect those that cannot be

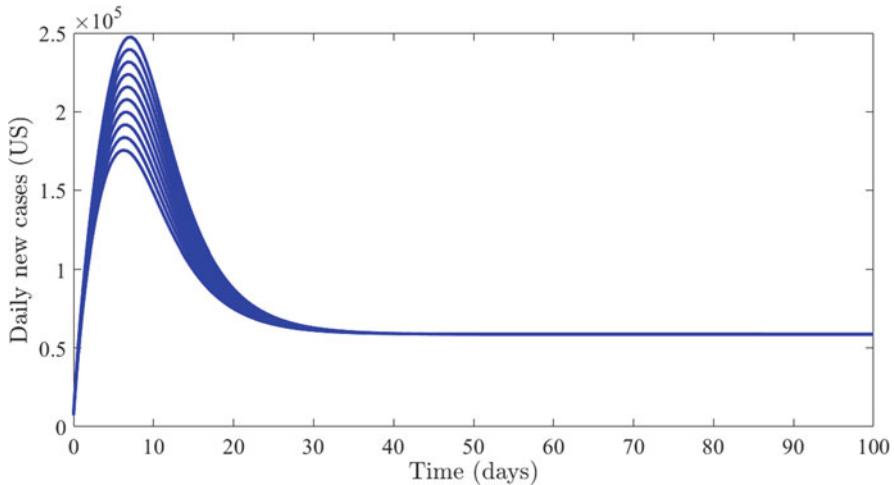


Fig. 10.6 Simulations of the special case of the vaccination model (10.1), for the number of daily new cases in the United States as a function of time, showing convergence of initial conditions to the unique endemic equilibrium when $\tilde{\mathbb{R}}_v > 1$. The values of the parameters used in these simulations are as given by their baseline values given in Tables 10.3 and 10.4, with $\Pi = 12000 \text{ day}^{-1}$, $\beta_p = 0.9909 \text{ day}^{-1}$, $\beta_s = 0.9986 \text{ day}^{-1}$, $\beta_a = 0.9942 \text{ day}^{-1}$, $\beta_h = 0.9989 \text{ day}^{-1}$, and $\gamma_h = 0.02 \text{ day}^{-1}$ (so that, $\tilde{\mathbb{R}}_v = 2.7444 > 1$)

vaccinated (so that vaccine-induced herd immunity is achieved in the population). Specifically, we let $\frac{V^*}{N^*}$ be the proportion of vaccinated susceptible individuals at the disease-free steady state.

To compute the herd immunity threshold associated with the vaccination model (10.1), we set the vaccination reproduction number (\mathbb{R}_{cv} ; defined in Eq. (10.7)) to one and solve for f_v . This gives:

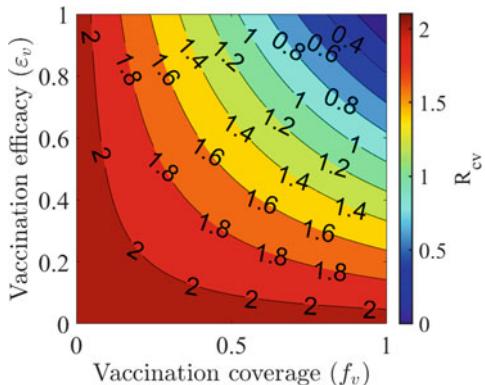
$$f_v = \frac{1}{\varepsilon_v} \left[\left(1 - \frac{1}{\mathbb{R}_0} \right) \right] = f_v^c \quad (\text{for } \mathbb{R}_0 > 1). \quad (10.18)$$

It follows from Eq. (10.18) that $\mathbb{R}_{cv} < (>) 1$ if $f_v > (<) f_v^c$. Furthermore, $\mathbb{R}_{cv} = 1$ whenever $f_v = f_v^c$. This result is summarized below:

Theorem 8 *Vaccine-induced herd immunity (i.e., COVID-19 elimination) can be achieved in the United States, using any of the FDA-approved anti-COVID vaccines, if the vaccine-derived immunity resulted in $f_v > f_v^c$ (i.e., if $\mathbb{R}_{cv} < 1$). If $f_v < f_v^c$ (i.e., if $\mathbb{R}_{cv} > 1$), then the vaccination program will fail to eliminate the pandemic.*

Epidemiologically speaking, Theorem 8 implies that the use of any of the approved COVID-19 vaccines can lead to the elimination of the pandemic in the United States if the proportion of susceptible individuals fully vaccinated at steady state reached or exceeded the aforementioned critical threshold value. In other words, the SARS-CoV-2 pandemic will be eliminated in the United States if $f_v > f_v^c$.

Fig. 10.7 Contour plot of the vaccine reproduction number (\mathbb{R}_{cv}) of the model (10.1), as a function of vaccine coverage (f_v) at steady state and vaccine efficacy (ε_v), for the United States. The values of the parameters used in these simulations are as given by their baseline values given in Tables 10.3 and 10.4



On the contrary, the vaccination program will fail to eliminate the pandemic if the proportion of the vaccinated individuals at the disease-free equilibrium falls below the aforementioned critical herd immunity threshold.

Figure 10.7 depicts a contour plot of the *vaccine reproduction number* (\mathbb{R}_{cv}), as a function of vaccine efficacy (ε_v) and vaccine coverage (f_v) at the disease-free steady state. For these simulations, the value of each of the parameters of the model is maintained at baseline as tabulated in Tables 10.3 and 10.4. As expected, this figure shows a decrease in the value of \mathbb{R}_{cv} with increasing efficacy and coverage of the vaccine. Furthermore, the contour plot shows that for the case where the overall average protective efficacy of the three vaccines is set at 85% (as tabulated in Table 10.3), at least 62% of the wholly susceptible population needs to be fully vaccinated at steady state to bring the *vaccination reproduction number* (\mathbb{R}_{cv}) below one (Fig. 10.7). However, if the average vaccine efficacy of the three vaccines drops to 60% (which is plausible, since data shows that the efficacy of the three vaccines against Omicron is much lower than against other variants [10, 66, 72, 78]), then the requirement for achieving the vaccine-derived herd immunity threshold in the United States drastically increases to 87%. In other words, based on the results depicted in Fig. 10.7, this study shows that the prospects of achieving vaccine-derived herd immunity in the United States, using the aforementioned three vaccines (Pfizer, Moderna, and Johnson & Johnson vaccine), are promising provided the average vaccine efficacy against the predominant Omicron variant is high enough (even if the vaccine coverage is moderate).

10.3.4 Global Parameter Sensitivity Analysis

The vaccination model (10.1) contains 22 parameters. Although baseline values of these parameters are given mostly based on published study (as tabulated in Tables 10.3 and 10.4), uncertainties are expected to arise in the estimate of these parameter values. It is, therefore, crucial to assess the impact of these uncertainties

on the outcome of the model simulations. It is also important to determine which of the 22 parameters have the most influence on the dynamics of the model (with respect to a certain chosen response function). In this section, we will use Latin hypercube sampling technique and partial rank correlation coefficients (PRCCs) to quantify those parameters that have the highest impact on the value of the chosen response function [6, 52, 54].

Parameter sensitivity analysis is a quantitative measure for determining the extent to which a chosen response function changes with respect to variations in the input variables (i.e., parameters of the model) [30, 51, 52]. For the purpose of this study, the *vaccination reproduction number* of the model (\mathbb{R}_{cv}) is chosen as the response function. It should be mentioned that, since the values of 4 of the 22 parameters of the vaccination model (namely, the demographic parameters Π and μ , and the disease-induced mortality rates of pre-symptomatic and asymptomatic individuals, δ_p and δ_a) are reliably known (from census data and due to the assumptions we made in Sect. 10.2 about the values of δ_a and δ_p), they are excluded from the sensitivity analysis. In other words, the sensitivity analysis will be based on the remaining 18 parameters of the vaccination model (10.1).

The process of carrying out the sensitivity analysis entails defining a range (lower and upper bound) and distribution for each parameter of the model and then splitting each parameter range into 1,000 equal sub-intervals [30, 51]. In this study, the range for each of the parameters of the model considered in the sensitivity analysis is obtained by taking 20% to the left and right of its respective baseline value given in Table 10.6 [30]. Furthermore, it is assumed, for statistical tractability, that all parameters in the response function obey the uniform distribution [30, 51]. Sets of parameter values are sampled (or drawn) from this space (i.e., parameter ranges) without replacement and used to form a 1000×18 matrix (or hypercube). Each row of this matrix is used to compute the response function (\mathbb{R}_{cv}) and the PRCC values are then computed to assess the contributions of uncertainty and variability in individual parameters to uncertainty and variability in the *vaccine reproduction number* [30]. Parameters with high PRCC values close to -1 or $+1$ are said to be highly correlated with the response function [30, 52]. Those with negative (positive) PRCC values are said to be negatively (positively) correlated with the response function (\mathbb{R}_{cv}) [30].

Figure 10.8 shows the PRCC values of the 18 parameters of the vaccination model (10.1), with respect to the chosen response function (\mathbb{R}_{cv}), computed on Day 43 of the onset of the Omicron variant in the United States (i.e., the PRCC values were computed on January 9, 2022). This date was chosen (to compute a snapshot of the PRCC values) because it corresponds to the time when the model predicted a peak of the daily new cases (see the blue curve in Fig. 10.2). The PRCC values are also tabulated in Table 10.6. It can be seen from Fig. 10.8 that the top five parameters that have the most influence on the response function (\mathbb{R}_{cv}) are:

- (i) The effective contact rate for pre-symptomatically infectious individuals (β_p).
- (ii) The effective contact rate for asymptomatically infectious individuals (β_a).
- (iii) Progression rate of pre-symptomatic individuals (σ_p).

Table 10.6 Table of PRCC values of the parameters in the expression for the vaccination reproduction number, \mathbb{R}_{cv} , of the vaccination model (10.1). PRCC values above 0.5 in magnitude are highlighted with a *, implying that these parameters are highly correlated with the response function (i.e., they significantly impact the value of the response function, \mathbb{R}_{cv}). Apart from the efficacies (i.e., ε_v , ε_n , and ε_{nv}) and the proportion “ r ”, which are dimensionless, all the other parameters and their ranges have unit of *per day*

Parameter	Baseline value	Range	PRCC: \mathbb{R}_{cv}
β_p	0.2309 day^{-1}	$0.18472\text{--}0.27708$	0.815*
β_s	$9.998 \times 10^{-4} \text{ day}^{-1}$	$0.00079\text{--}0.00119$	0.0153
β_a	0.5429 day^{-1}	$0.43436\text{--}0.65155$	0.981*
β_h	$4.998 \times 10^{-5} \text{ day}^{-1}$	$3.9 \times 10^{-5}\text{--}5.9 \times 10^{-5}$	-0.022
σ_E	0.2000 day^{-1}	$0.16000\text{--}0.24000$	0.0094
σ_p	0.5000 day^{-1}	$0.40000\text{--}0.60000$	-0.820*
r	0.095 (dimensionless)	$0.07600\text{--}0.11400$	-0.467
γ_s	0.1000 day^{-1}	$0.08000\text{--}0.12000$	-0.852*
γ_a	0.2000 day^{-1}	$0.16000\text{--}0.24000$	0.0240
γ_h	0.1250 day^{-1}	$0.10000\text{--}0.15000$	-0.035
ω_v	0.0037 day^{-1}	$0.00291\text{--}0.00437$	0.0252
δ_s	$4.980 \times 10^{-5} \text{ day}^{-1}$	$3.9 \times 10^{-5}\text{--}5.9 \times 10^{-5}$	0.0051
δ_h	$5.0 \times 10^{-5} \text{ day}^{-1}$	$4.0 \times 10^{-5}\text{--}6.0 \times 10^{-5}$	0.0016
ξ_v	$1.9 \times 10^{-5} \text{ day}^{-1}$	$1.5 \times 10^{-5}\text{--}2.2 \times 10^{-5}$	0.0105
ϕ_s	0.2000 day^{-1}	$0.16000\text{--}0.24000$	-0.956*
ε_v	0.8500 (dimensionless)	$0.68000\text{--}1.02000$	-0.0478
ε_n	0.8500 (dimensionless)	$0.68000\text{--}1.02000$	-0.0285
ε_{nv}	0.9500 (dimensionless)	$0.76000\text{--}1.14000$	-0.0153

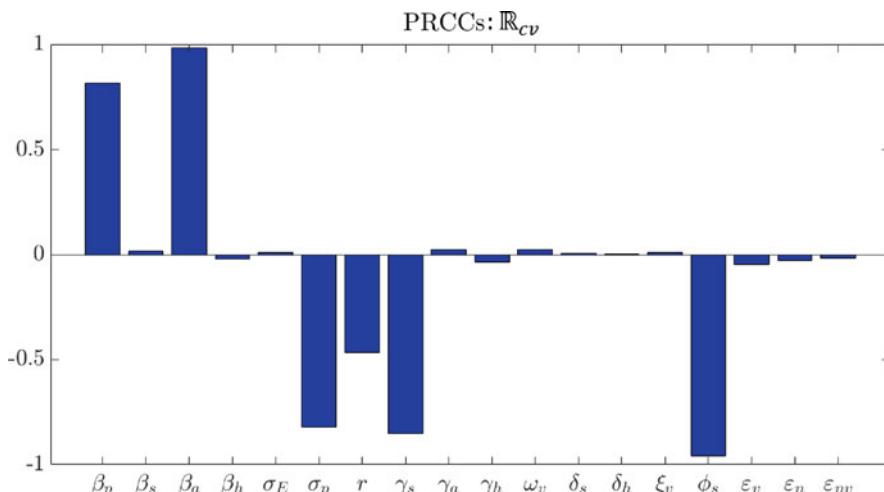


Fig. 10.8 Partial rank correlation coefficients (PRCCs) of the parameters of the vaccination model (10.1) with respect to the response function (\mathbb{R}_{cv}). Snapshot of the PRCC plot generated on Day 43 of the onset of the Omicron variant in the United States (i.e., when the model predicted a peak of the daily new cases, as shown by the blue curve in Fig. 10.2). Parameter values used in these simulations are as given by the baseline values and their corresponding ranges, tabulated in Table 10.6

- (iv) Recovery rate of symptomatic individuals (γ_s).
- (v) Hospitalization rate of the individuals with clinical symptoms (ϕ_s).

Hence, it follows from the parameter sensitivity analysis that implementing public health intervention and mitigation strategies that target reducing the effective contact rates for pre-symptomatic and asymptomatic individuals (β_p and β_a) will be very effective in reducing the response function \mathbb{R}_{cv} (and, consequently, reduce the burden of the pandemic in the United States). The parameters β_p and β_a can be reduced by implementing control strategies such as social distancing, community lockdowns, use of face masks in public, quarantine of suspected cases, and isolation of those with clinical symptoms of the disease. Furthermore, strategies that increase the progression rate of pre-symptomatic individuals (σ_p) and recovery rates of the individuals in the symptomatic class (γ_s) as well as increase the detection and hospitalization of symptomatic cases (ϕ_s) will reduce the COVID-19 burden in the community. The progression rate of pre-symptomatic individuals (σ_p) can be increased by contact tracing of confirmed SARS-CoV-2 cases. The parameters γ_s and ϕ_s can be increased by implementing control strategies such as the treatment of confirmed SARS-CoV-2 cases and the rapid detection (*via* the use of effective and large-scale random diagnostic testing) and hospitalization of symptomatic cases.

In conclusion, this study identifies five parameters (β_p , β_a , σ_p , γ_a , and ϕ_s) that have the greatest influence on the value of the vaccination reproduction number (\mathbb{R}_{cv}), which governs the persistence or effective control of the pandemic in the United States. It is worth stating that this result is consistent with some of the results reported in the SARS-CoV-2 modeling literature, such as those in [30, 56, 59] which suggest that pre-symptomatic and asymptomatic individuals are the main drivers of the COVID-19 pandemic. Hence, to effectively control the SARS-CoV-2 pandemic, the public health control and mitigation strategies should be focused on effectively targeting the five identified parameters.

10.4 Numerical Simulations

Having rigorously analyzed the qualitative dynamics of the vaccination model (10.1) and carrying out the detailed global sensitivity analysis of its parameters, the vaccination model will now be numerically simulated to assess the population-level impact of control and mitigation strategies against the SARS-CoV-2 pandemic in the United States. The main focus of these simulations is to assess the impacts of face mask usage in public (as a singular intervention) and the combined impact of face mask usage with vaccination (using any of the three vaccines, Pfizer, Moderna, and Johnson & Johnson, being administered in the United States) on limiting or curtailing the burden of the COVID-19 pandemic in the United States. Unless otherwise stated, the simulations of the vaccination model (10.1) will be carried out using the baseline values of the fixed and fitted parameters given in Tables 10.3 and 10.4.

10.4.1 Effect of Masking as a Singular Control and Mitigation Intervention

The use of face masks in public has played a very significant dual role of preventing people from getting infected with COVID-19 (*inward efficacy* or respiratory protection), in addition to preventing those infected with COVID-19 from infecting others (*source control*) [18]. As currently formulated, the vaccination model (10.1) does not explicitly account for the impact of face mask usage in public. In order to incorporate the usage of masking into the vaccination model, we re-scale the community contact rate parameters (β_p , β_s , β_a , and β_h) by a measure of face masks effectiveness in prevention acquisition or transmission of infection. In particular, we carried out the following parameter re-scaling (where the symbol \rightarrow means “replaced by”) in the model (10.1):

$$\begin{aligned}\beta_p &\rightarrow \beta_p(1 - \varepsilon_m c_m), \quad \beta_s \rightarrow \beta_s(1 - \varepsilon_m c_m), \quad \beta_a \rightarrow \beta_a(1 - \varepsilon_m c_m) \quad \text{and} \\ \beta_a &\rightarrow \beta_h(1 - \varepsilon_m c_m),\end{aligned}\tag{10.19}$$

where $0 \leq \varepsilon_m \leq 1$ is the efficacy of the face mask to prevent transmission or acquisition of infection and $0 \leq c_m \leq 1$ is the compliance in face mask usage in the community. For the aforementioned masking scenario, the associated *masking reproduction number*, denoted by \mathbb{R}_{cm} , is given by (where \mathbb{R}_{cv} is as defined in Eq. (10.7), but with the re-scaling (10.19) used in place of the infection rates):

$$\mathbb{R}_{cm} = (1 - \varepsilon_m c_m) \mathbb{R}_{cv}.\tag{10.20}$$

The simulation results obtained, shown by the contour plots depicted in Fig. 10.9, show a marked decrease in the masking reproduction number (\mathbb{R}_{cm}) with increasing mask efficacy and compliance in mask usage. Using moderately efficacious face mask in the community, such as a face mask with efficacy 70% (e.g., the surgical mask with proper fitting), the simulation results show that at least 75% of the populace need to be consistently wearing face mask in public to reduce (and maintain) the masking reproduction number (\mathbb{R}_{cm}) to a value less than unity (it is worth recalling that bringing the masking reproduction number of the re-scaled version of the vaccination model (10.1) to a value less than unity is a necessary and sufficient condition for the elimination of the disease, in line with Theorems 2, 4, and 5 for the global asymptotic stability of the disease-free equilibrium of the model). In other words, this study shows that the use of face mask as a singular intervention can lead to the effective control (or elimination) of the SARS-CoV-2 pandemic if at least 75% of the populace consistently wear a face mask of moderate efficacy (e.g., the surgical mask). If masks of higher efficacy (e.g., N-95 mask or equivalent) are favored instead, our simulations show that such elimination can be achieved if 55% of the populace consistently wear these masks in public. Hence, the

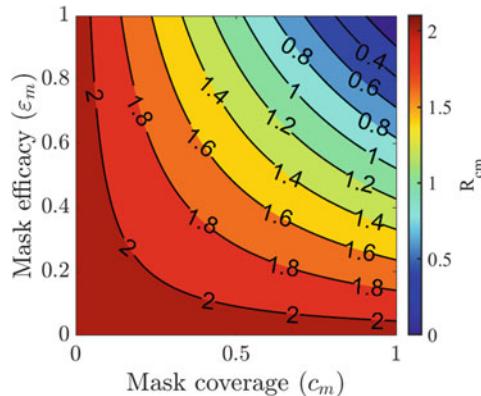
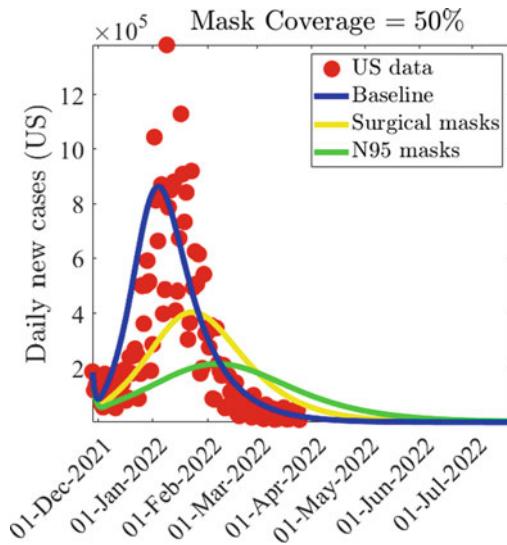


Fig. 10.9 Simulations of the re-scaled version of the vaccination model (10.1) to assess the effect of face mask usage as a singular public health control and mitigation intervention (i.e., no vaccination). Contour plots of the masking reproduction number (R_{cm}) of the re-scaled version of the model (10.1), as a function of mask coverage (c_m) and efficacy (e_m), for the United States. The other parameter values used in these simulations are as given by their respective baseline values in Tables 10.3 and 10.4, with the vaccine-related parameters and state variable (ε_v , ω_v , and $V(t)$) set to zero

community-wide masking coverage needed to eliminate the disease decreasing with increasing efficacy of the face mask type favored or prioritized in the community.

We further simulated the re-scaled version of the vaccination model (10.1) (given by (10.1) with the re-scaling (10.19)) to assess the impact of face mask usage as a singular public health control and mitigation intervention (i.e., in the absence of vaccination) on the daily new cases in the United States. To assess the singular impact of face mask usage, these simulations are carried out for the special case of the re-scaled version of the model with no vaccination (i.e., all the vaccine-related parameters and state variable of the re-scaled model are set to zero). Furthermore, for these simulations, we set the face mask coverage in the community to be 50% (i.e., $c_m = 0.5$) and consider the scenario where only two mask types, namely, surgical (with estimated efficacy of 70%, so that $\varepsilon_v = 0.7$) and N-95 masks or equivalent (with estimated efficacy of 95%, so that $\varepsilon_v = 0.95$), are prioritized in the community. The simulation results for the two mask types, depicted in Fig. 10.10, show a significant decrease in the average daily new cases at the peak recorded under the baseline scenario (i.e., compare the peaks for the yellow and green curves with the peak of the blue curve, which represents the baseline scenario; note that, for the data used in these simulations, the peak for the baseline scenario occurred on January 3, 2022). For instance, under this scenario (in the absence of vaccination), if surgical masks are prioritized (and with 50% coverage), about 53% of the daily cases recorded at the peak under the baseline scenario will have been prevented (Fig. 10.10, compare the peaks of the blue and gold curves). Furthermore, if N-95 masks or equivalent are prioritized (with the same 50% coverage), about 75% of

Fig. 10.10 Simulations of the re-scaled version of the vaccination model, given by (10.1) with the re-scaling (10.19), to assess the effect of face mask usage as a singular public health control and mitigation intervention (i.e., no vaccination) on the average daily new cases of SARS-CoV-2 in the United States. The other parameter values used in these simulations are as given by their respective baseline values in Tables 10.3 and 10.4, with the vaccine-related parameters and state variable (ε_v , ω_v , and $V(t)$) set to zero



the daily new cases recorded at the peak of the baseline scenario would have been averted (compare peaks of green and blue curves in Fig. 10.10).

In summary, this study shows that the prospect of eliminating the SARS-CoV-2 pandemic using masking as a singular public health control and mitigation strategy is promising, provided masks of moderate or high efficacy (with moderate to high coverage) are prioritized. Specifically, this study showed, based on the current observed data used in our simulations (i.e., based on the SARS-CoV-2 case data for the period November 28, 2021, to March 23, 2022, used to parameterize the model), i.e., from the contour plot (Fig. 10.9), that the SARS-CoV-2 pandemic can be eliminated in the United States (i.e., suppressed from taking off) if approximately half the populace were consistently wearing N-95 mask (or equivalent) in public (if surgical masks were prioritized, the coverage level needed to achieve such elimination increases to 75%). Although it may not be practical to expect humans to always be wearing face masks in public, implementing masking as a singular strategy is important for many reasons, including: (a) disease burden (i.e., severe disease, hospitalization, and death) which can be significantly reduced and (b) buying time (by suppressing the burden of the disease and saving lives) before a safe and effective vaccine becomes available. For this data set, if masking was started on Day 1 of the onset of Omicron (i.e., November 28, 2021), this study showed that up to 53–75% of the daily new cases recorded at the peak (this corresponds to 225000–425200) will have been prevented.

10.4.2 Assessing the Combined Impact of Vaccination and Masks on Herd Immunity Threshold

The re-scaled version of the vaccination model (10.1) is simulated (using the baseline values of the fixed, fitted, and assumed parameter values in Tables 10.3 and 10.4), to assess the combined impact of vaccination (at baseline) and masking, on the dynamics of COVID-19 in the United States. Figure 10.11 depicts contour plots of the masking reproduction number (\mathbb{R}_{cm}) of the modified version of model (10.1), as a function of the average vaccine efficacy (ε_v) of the three vaccines (namely, Pfizer, Moderna, and Johnson & Johnson vaccine) against the acquisition of infection with Omicron and the fraction of the US population fully vaccinated at steady state (f_v). Numerical simulations are carried out for the scenario when the baseline face mask usage in the community is increased by 20%, for various face mask types. Figure 10.11a shows that if fabric or cloth masks (with low masking efficacy, i.e., $\varepsilon_m = 0.30$) are prioritized, the herd immunity requirement corresponding to 62% now reduces to 58%. However, if moderately effective procedure or surgical masks (with $\varepsilon_m = 0.70$) are prioritized, the herd immunity requirement significantly reduces to 52% (Fig. 10.11b). Furthermore, the vaccine-derived herd immunity threshold reduces drastically from 62% to 48% if highly effective N-95 masks (with $\varepsilon_m = 0.95$) are prioritized (Fig. 10.11c). In summary, the contour plots in Fig. 10.11a–c show that the proportion of the individuals who need to be fully vaccinated to achieve herd immunity in the United States reduces with increasing coverage of face masks in the community (from the baseline face mask usage). Furthermore, the level of reduction achieved depends on the quality of the face masks used (specifically, greater reduction in herd immunity level needed to eliminate the disease is achieved if the high-quality N-95 masks are prioritized, in

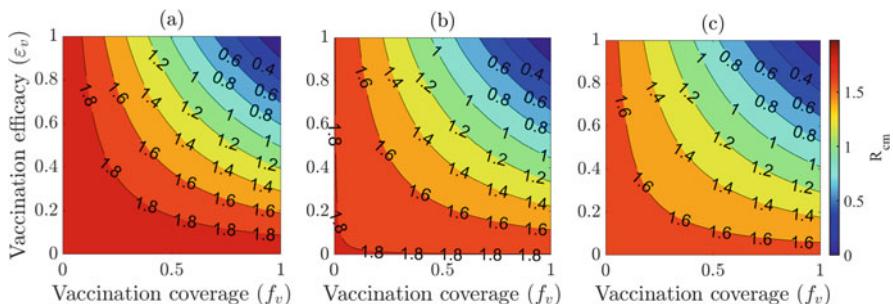


Fig. 10.11 (a–c) Contour plots of the face mask reproduction number (\mathbb{R}_{cm}), as a function of vaccine efficacy (ε_v) and the fraction vaccinated at the steady state (f_v) for the case when (a) cloth mask is prioritized and the mask coverage is increased by 20% from the baseline, (b) surgical mask is prioritized and the coverage in its usage is increased by 20% from the baseline, and (c) N-95 mask is prioritized and the coverage in its usage is increased by 20% from the baseline. The values of all other parameters used in the simulations are as given by the baseline values in Tables 10.3 and 10.4.

comparison to the scenario where the cloth masks or moderately effective surgical masks are prioritized).

10.4.3 Assessing the Combined Impact of Vaccination and Masks on Daily New Cases

The re-scaled version of the vaccination model (10.1) is further simulated (using the baseline values of the fixed, fitted, and assumed parameter values in Tables 10.3 and 10.4) to assess the combined impacts of mask coverage (c_m), mask type (cloth masks with masking efficacy of 30%, i.e., $\varepsilon_m = 0.30$; surgical masks with masking efficacy of 70%, i.e., $\varepsilon_m = 0.70$; and N-95 respirators with masking efficacy of 95%, i.e., $\varepsilon_m = 0.95$), and vaccination (with average vaccine efficacy at the baseline level) on the daily number of new COVID-19 cases in the United States. For the scenario when the mask coverage increases by 10% from its baseline value, the results obtained depicted in Fig. 10.12a show that using the ineffective cloth masks will result to a 4.08% reduction in the average number of new daily cases at the peak from the baseline (by comparing the magenta and blue curves in Fig. 10.12a). The reduction in the average number of peak daily new cases is more significant if face masks of higher quality are prioritized. Specifically, if the moderately effective surgical masks are prioritized, the simulation show that up to 9.79% reduction in peak level of the daily new cases at the peak, in comparison to the peak baseline

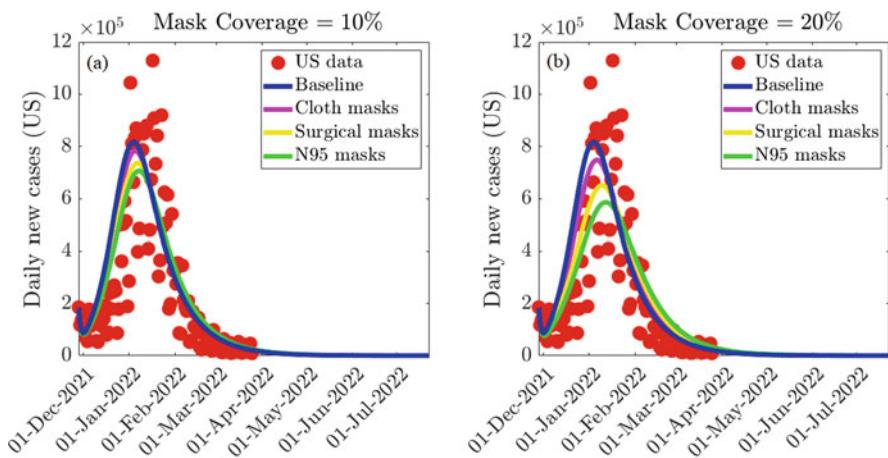


Fig. 10.12 Simulations of the model (10.1), showing the incremental impact of mask coverage (c_m) and mask type (cloth masks, with $\varepsilon_m = 0.3$; surgical masks, with $\varepsilon_m = 0.7$; and N-95 respirators, with $\varepsilon_m = 0.95$) on the daily new COVID-19 cases in the United States, as a function of time. In Fig. 10.12a, the mask coverage (c_m) is increased by 10% and in Fig. 10.12b, the mask coverage (c_m) is increased by 20%, from their respective baseline values. The values of the other parameters used in these simulations are as given in Tables 10.3 and 10.4

level (by comparing the gold curve and blue curve in Fig. 10.12a). The reduction increases significantly to 13.49% if the highly effective N-95 masks are prioritized (compare the blue and green curves in Fig. 10.12a). These reductions are more dramatic for the scenario when the mask coverage increases by 20% from its baseline value. The results obtained, depicted in Fig. 10.12b, show that using the ineffective cloth masks will result to a 8.35% reduction in the average daily new cases at the peak, in comparison to the baseline (compare the magenta and blue curves in Fig. 10.12b). The reduction in the average daily new cases at the peak is more notable if moderately effective surgical masks are prioritized; the simulation show that up to 20.37% reduction in peak daily new cases can be achieved, in comparison to the baseline (compare the gold curve and blue curve in Fig. 10.12b). This reduction in the average daily new cases at the peak is more drastic, about 28.25% (in comparison to the baseline) if the highly effective N-95 masks are prioritized (compare the green and blue curves in Fig. 10.12b). In summary, the simulations in Fig. 10.12a–b show that the reduction in the average number of daily new cases at the peak (from the baseline) is significant if the face mask coverage in the community is increased (from the baseline coverage level). Furthermore, the reduction in the average number of peak daily cases also depends on the quality of the face masks used (specifically, drastic reduction in peak daily new cases is noted if the high-quality N-95 masks are prioritized, in comparison to the scenario where the cloth masks or moderately effective surgical masks are prioritized).

10.5 Discussion and Conclusions

This chapter is based on the use of mathematical modeling approaches, coupled with rigorous analysis and computation, to address the problem of the spread and control of the novel 2019 coronavirus pandemic (COVID-19) in the United States. Specifically, a mathematical model, which takes the form of a deterministic system of nonlinear differential equations, is developed and used to assess the population-level impact of three of the four FDA-approved anti-COVID vaccines (Pfizer, Moderna, and Johnson & Johnson vaccines) on the transmission dynamics and control of the COVID-19 pandemic in the United States. The impact of face mask use strategy, implemented as a singular intervention strategy or in combination with the vaccination program, is also assessed.

The model was rigorously analyzed (using techniques, tools, and theories from nonlinear dynamical systems) to study its qualitative dynamical features. Furthermore, the model was parameterized by fitting it to the observed new daily COVID-19 case data for the United States for the period from November 28, 2021, to March 23, 2023 (this period was chosen to coincide with the time the Omicron variant first emerged in the United States). We specifically fitted the model using the segment of the data from November 28, 2021, to February 23, 2022, and use the remaining segment of the data (i.e., the segment from February 24, 2022, to March 23, 2022) to cross-validate the model. The cross-validation showed that the

vaccination model predicts the case data for the time period from February 24, 2022, to March 23, 2022, reasonably well.

The rigorous qualitative analysis of the vaccination model revealed that its unique disease-free equilibrium is locally asymptotically stable whenever the associated *control reproduction number* of the model (denoted by \mathbb{R}_{cv}) is less than one. It was also shown, using the theory of center manifold [8, 45], that the model undergoes the phenomenon of *backward bifurcation* when the control reproduction number of the model is less than one under certain conditions. The epidemiological implication of the backward bifurcation phenomenon is that the usual epidemiological requirement of having the control reproduction number of the model being less than 1, while necessary, is no longer sufficient for the elimination of the disease. In the presence of a backward bifurcation situation, more control resources need to be invested to further reduce the control reproduction number. We identified two main sufficient conditions for the presence of backward bifurcation in the vaccination model we presented, namely, (a) imperfect vaccine-derived, natural, and combined vaccine-derived and natural immunity to protect against the acquisition of infection (of vaccinated susceptible individuals) and reinfection (of recovered vaccinated or unvaccinated individuals) and (b) disease-induced mortality and reinfection of recovered individuals. In other words, the phenomenon of backward bifurcation does not occur when (i) the vaccine-derived and natural immunity (including the combined vaccine-derived and natural immunity) is perfect (i.e., when $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$) and (ii) when the disease-induced mortality is negligible (i.e., $\delta_p = \delta_s = \delta_a = \delta_h = 0$) and recovered individuals do not acquire reinfection (i.e., $\varepsilon_n = \varepsilon_{nv} = 1$). In any of these two scenarios where backward bifurcation does not occur, we proved the global asymptotic stability of the disease-free equilibrium when the associated control reproduction number is less than one (in this case, the model undergoes a forward bifurcation at the bifurcation point, where the associated control reproduction number equals one). The epidemiological implication of the global asymptotic stability result for the disease-free equilibrium of the model (for the aforementioned special cases of the model where backward bifurcation does not occur) is that the SARS-CoV-2 pandemic can be eliminated in the United States if the associated control reproduction number can be brought to (and maintained at) a value less than one (in other words, in this case, bringing and maintaining the associated control reproduction number to a value less than one is necessary and sufficient for the elimination of the pandemic in the United States).

We also showed that the model has a unique and locally asymptotically stable endemic equilibrium for a special case (where the vaccine is assumed to offer perfect protective efficacy against the acquisition of infection, no reinfection, no waning of vaccine-derived immunity, and disease-induced mortality is negligible) whenever the associated control reproduction number exceeds one. The epidemiological implication of this result is that, for this special case of the model, the disease will persist in the population whenever the associated reproduction number is greater than one. An explicit expression for the vaccine-induced herd immunity threshold for the United States was derived. It was shown, using current data for new daily COVID-19 cases in the United States, that, for the case where the three FDA-

approved vaccines offer average protective efficacy against the Omicron variant of about 85%, vaccine-derived herd immunity will be achieved in the United States if at least 62% of the populace is fully vaccinated. However, if the average cross-protection efficacy provided by the vaccines reduces slightly (e.g., to 60%), about 87% of the population needs to be fully vaccinated with either of the aforementioned vaccines to achieve the vaccine-derived herd immunity.

Furthermore, using global sensitivity analysis, we identified the parameters of the model that have the most influence on the control reproduction number of the model, \mathbb{R}_{cv} (hence disease burden in the community). Specifically, we identified the top five PRCC-ranked parameters to be the effective contact rate for pre-symptomatically infectious individuals (β_p), asymptomatically infectious individuals (β_a), the progression rate of pre-symptomatic individuals (σ_p), the recovery rate of symptomatic individuals (γ_s), and the hospitalization rate of the individuals with clinical symptoms (ϕ_s). The numerical PRCC values indicate that reduction of the infection rate of pre-symptomatic and asymptomatic individuals results in the reduction of \mathbb{R}_{cv} . The parameters β_p and β_a can be reduced by implementing control strategies, such as social distancing, community lockdowns, use of face masks in public, quarantine, and isolation of the confirmed cases of COVID-19. Furthermore, by increasing the progression rate of pre-symptomatic individuals, increasing the recovery rate of symptomatic individuals, and the increase in the detection and hospitalization of symptomatic cases will ultimately reduce \mathbb{R}_{cv} . The progression rate of the pre-symptomatic individuals (σ_p) can be increased by implementing non-pharmaceutical control invention, such as contact tracing. The parameters γ_s and ϕ_s can be increased by implementing control strategies, such as treatment of the COVID-19-infected individuals and the detection and hospitalization of symptomatic cases by large-scale blanket testing in the community.

The model was adapted and used to assess the population-level impact of using face mask in the community as a singular intervention strategy (i.e., in the absence of vaccination) and also in combination with the vaccination program. We assessed exclusively the impact of face masks as a singular public health control and mitigation intervention (i.e., with the vaccine-related parameters and state variable set to zero) on the herd immunity requirement and on the average daily new cases in the United States. Under this scenario (i.e., with no vaccination), we showed that the prospect of COVID-19 elimination in the United States is enhanced if almost half of the populace is consistently wearing N-95 mask (if surgical masks were prioritized, the coverage level needed to achieve such elimination increases to 75%). Further, we simulated the model for the same scenario (i.e., in the absence of vaccination) to assess the impact of masking on the peak daily new cases. Our simulations showed a significant decrease in the peak daily new cases (i.e., up to 53%) if the implementation of moderately effective surgical masks (with 50% coverage) was started on Day 1 of the onset of Omicron (i.e., end of November 2021). Marked reductions in the peak daily new cases were recorded (i.e., up to 75%) if N-95 masks were prioritized for the same scenario (i.e., if masking was started on Day 1 of the onset of Omicron and its coverage is increased by 50% from the baseline).

The re-scaled version of the vaccination model was also used to assess the population-level impact of vaccination (at the baseline) combined with face mask usage for various mask types. For the scenario when the baseline face mask usage in the community is increased by 20% and moderately effective surgical mask are prioritized, the simulations showed that at least 52% of the populace need to be fully vaccinated in order to achieve vaccine-derived herd immunity. However, if highly efficacious face masks are prioritized (such as N-95 mask), our simulations showed that the requirement for achieving the vaccine-derived herd immunity is reduced significantly to 48%. Furthermore, for the same scenario (i.e., baseline face mask usage is increased by 20%), the simulations show that if the highly efficacious N-95 masks are prioritized then the average number of daily new cases at the peak is significantly reduced from the baseline by about 28%. In summary, the theoretical and numerical simulation results generated from this chapter showed that the prospects for the effective control and elimination of the COVID-19 pandemic in the United States are significantly improved if vaccination is combined with a face mask strategy (that prioritizes moderately effective and high-quality masks), particularly if the average efficacy of the three of the four FDA-approved vaccines being administered in the United States (namely, Pfizer, Moderna, and Johnson & Johnson vaccine) is high (i.e., $\approx 85\%$).

It should be mentioned that in this chapter, we used a relatively basic model for vaccination against the SARS-CoV-2 pandemic to illustrate the epidemiological concepts and features we are highlighting. The model can be extended to incorporate other important features associated with the vaccination and SARS-CoV-2 immunity, such as allowing for waning of natural and the combined natural and vaccine-derived immunity, boosting of immunity (especially vaccine-derived) [61, 66], and the effect of human behavior changes with respect to control and mitigation interventions (particularly adherence to vaccination and/or face mask usage) [60].

Acknowledgments ABG acknowledges the support, in part, of the National Science Foundation (grant number: DMS-2052363). SS acknowledges the support of the Fulbright Foreign Student Program.

Appendix 1: Proof of Theorem 3

Proof The proof is based on the center manifold theory [8, 82], and to apply this theory, it is convenient to introduce the following change of variables: let $S = x_1$, $V = x_2$, $E = x_3$, $I_p = x_4$, $I_s = x_5$, $I_a = x_6$, $I_h = x_7$, $R_n = x_8$, $R_{nv} = x_9$. Using this transformation, the vaccination model (10.1) can be re-written in general form $\frac{dX}{dt} = (f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9)^T$, with $X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)^T$. Specifically, the model (10.1) can be written in terms of the transformed variables as:

$$\left\{ \begin{array}{l} \frac{dx_1}{dt} = f_1 = \Pi + \omega_v x_2 - \lambda x_1 - k_1 x_1, \\ \frac{dx_2}{dt} = f_2 = \xi_v x_1 - (1 - \varepsilon_v) \lambda x_2 - k_2 x_2, \\ \frac{dx_3}{dt} = f_3 = \lambda x_1 + (1 - \varepsilon_v) \lambda x_2 + (1 - \varepsilon_n) \lambda x_8 + (1 - \varepsilon_{nv}) \lambda x_9 - k_3 x_3, \\ \frac{dx_4}{dt} = f_4 = \sigma_e x_3 - k_4 x_4, \\ \frac{dx_5}{dt} = f_5 = r \sigma_p x_4 - k_5 x_5, \\ \frac{dx_6}{dt} = f_6 = (1 - r) \sigma_p x_4 - k_6 x_6, \\ \frac{dx_7}{dt} = f_7 = \phi_s x_5 - k_7 x_7, \\ \frac{dx_8}{dt} = f_8 = \gamma_s x_5 + \gamma_a x_6 + \gamma_h x_7 - (1 - \varepsilon_n) \lambda x_8 - k_8 x_8, \\ \frac{dx_9}{dt} = f_9 = \xi_v x_8 - (1 - \varepsilon_{nv}) \lambda x_9 - k_9 x_9, \end{array} \right. \quad (10.21)$$

with the force of infection under the aforementioned transformation now defined as:

$$\lambda = \frac{(\beta_p x_4 + \beta_s x_5 + \beta_a x_6 + \beta_h x_7)}{N},$$

where

$$\begin{aligned} N &= x_1 + x_2 + x_3 + x_4 + x_5 + x_5 + x_6 + x_7 + x_7 + x_8 + x_9, k_1 = \xi_v + \mu, \\ k_2 &= \omega_v + \mu, k_3 = \sigma_E + \mu, k_4 = \sigma_p + \delta_p + \mu, k_5 = \phi_s + \gamma_s + \delta_s + \mu, \\ k_6 &= \gamma_a + \delta_a + \mu, k_7 = \gamma_h + \delta_h + \mu, k_8 = \xi_v + \mu, \text{ and } k_9 = \mu. \end{aligned}$$

The Jacobian of the system (10.21), evaluated at the DFE (\mathcal{E}_0), is given by:

$$J(\mathcal{E}_0) = \begin{pmatrix} -k_1 & \omega_v & 0 & -\beta_p J_1 & -\beta_s J_1 & -\beta_a J_1 & -\beta_h J_1 & 0 & 0 \\ \xi_v & -k_2 & 0 & -\beta_p J_2 & -\beta_s J_2 & -\beta_a J_2 & -\beta_h J_2 & 0 & 0 \\ 0 & 0 & -k_3 & -\beta_p J_3 & -\beta_s J_3 & -\beta_a J_3 & -\beta_h J_3 & 0 & 0 \\ 0 & 0 & \sigma_e & -k_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & r \sigma_p & -k_5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & (1 - r) \sigma_p & 0 & -k_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi_s & 0 & -k_7 & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma_s & \gamma_a & \gamma_h & -k_8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi_v & -k_9 \end{pmatrix},$$

where $J_1 = \frac{\omega_v + \mu}{\omega_v + \xi_v + \mu}$, $J_2 = \frac{(1 - \varepsilon_v) \xi_v}{\omega_v + \xi_v + \mu}$, and $J_3 = J_2 - J_1$.

Choosing β_p as the bifurcation parameter, and solving for β_p from $\mathbb{R}_{cv} = 1$ (i.e., at the bifurcation point), gives:

$$\beta_p = \frac{1}{(1 - \varepsilon_v f_v) \left(\frac{\sigma_E}{\sigma_E + \mu} \right) \left(\frac{1}{\sigma_p + \delta_p + \mu} \right) (D_p)} = \beta_p^*,$$

where

$$D_p = \{1 + d_1 r \sigma_p B_1 + d_2 (1 - r) \sigma_p B_2 + d_3 r \sigma_p B_1 B_3\},$$

with,

$$\begin{aligned} B_1 &= \left(\frac{1}{\phi_s + \gamma_s + \delta_s + \mu} \right), \quad B_2 = \left(\frac{1}{\gamma_a + \delta_a + \mu} \right) \\ B_3 &= \left(\frac{1}{\gamma_h + \delta_h + \mu} \right), \quad d_1 = \frac{\beta_s}{\beta_p}, \\ d_2 &= \frac{\beta_a}{\beta_p} \text{ and } d_3 = \frac{\beta_h}{\beta_p}. \end{aligned}$$

Let $J_{\beta_p^*}$ denote the Jacobian of the system (10.21) evaluated at the DFE (\mathcal{E}_0). It can be seen that the system (10.21), with $\beta_p = \beta_p^*$, has a simple eigenvalue with zero real part and all other eigenvalues having negative real part [5]. Hence, the center manifold theory [8, 9] can be applied to analyze the dynamics of the vaccination model (10.1) near the bifurcation point (where $\beta_p = \beta_p^*$). To apply the theory (in particular, the approach in [9]), the following computations (associated with the left and right eigenvectors of $J_{\beta_p^*}$, corresponding to the zero eigenvalue) are necessary.

□

Computation of Left and Right Eigenvectors of $J_{\beta_p^*}$

It can be seen that the left eigenvector of $J_{\beta_p^*}$, corresponding to the zero eigenvalue, is given by:

$\mathbf{v} = [v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9]$, where (noting that J_1 , J_2 , and J_3 are defined above):

$$\begin{aligned} v_1 &> 0 \text{ (free)}, \quad v_3 > 0 \text{ (free)}, \quad v_8 = v_9 = 0, \\ v_2 &= \frac{\omega_v v_1}{k_2}, \quad v_4 = \frac{k_3 v_3}{\sigma_E}, \quad v_5 = \frac{\beta_s (-J_1 v_1 - J_2 v_2 + J_3 v_3)}{k_5}, \\ v_6 &= \frac{\beta_a (-J_1 v_1 - J_2 v_2 + J_3 v_3)}{k_6}, \quad v_7 = \frac{\beta_h (-J_1 v_1 - J_2 v_2 + J_3 v_3)}{k_7}. \end{aligned} \tag{10.22}$$

Furthermore, the right eigenvector of $J_{\beta_p^*}$, corresponding to the zero eigenvalue, is given by:

$\mathbf{w} = [w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9]^T$, where:

$$\begin{aligned} w_1 &> 0 \text{ (free)}, \quad w_2 = \frac{k_1 w_1 + J_1 (\beta_p w_4 + \beta_s w_5 + \beta_a w_6 + \beta_h w_7)}{\omega_v}, \\ w_3 &= \frac{k_4 w_4}{\sigma_E}, \quad w_4 = \frac{k_5 w_5}{r \sigma_p}, \quad w_5 = \frac{k_7 w_7}{\phi_s}, \\ w_6 &= \frac{(1-r) \sigma_p w_4}{k_6}, \quad w_7 > 0 \text{ (free)}, \quad w_8 = \frac{k_9 w_9}{\xi_v}, \quad w_9 > 0 \text{ (free)}. \end{aligned} \quad (10.23)$$

For computational convenience, we set, without loss of generality, the components of the left eigenvectors v_1 and v_3 in (10.22) to one. Similarly, we set the components w_1, w_7 , and w_9 , of the right eigenvector, given in (10.23) to unity.

Computation of Backward Bifurcation Coefficients, a and b

The local bifurcation analysis near the bifurcation point ($\beta_p = \beta_p^*$) is determined by the signs of two bifurcation coefficients, denoted by a and b [8, 9]. Following [9], the expressions for these bifurcation coefficients are, respectively, given by:

$$a = \sum_{k,i,j=1}^n v_k w_i w_j \frac{\partial^2 f_k}{\partial x_i \partial x_j}(\mathcal{E}_0, \beta_p^*), \quad (10.24)$$

and

$$b = \sum_{k,i=1}^n v_k w_i \frac{\partial^2 f_k}{\partial x_i \partial \beta_p}(\mathcal{E}_0, \beta_p^*). \quad (10.25)$$

It can be shown, by substituting the expressions for the eigenvectors (w_k and v_k ; $k = 1, \dots, 9$) given in (10.22) and (10.23) and the partial derivatives of the functions f_k ($k = 1, \dots, 9$) defined in (10.21) into the expressions (10.24) and (10.25), that the bifurcation coefficients now become:

$$a = \frac{1}{D_a} \left[(C_a) \left\{ (P + Q + R) - (S + T + U) \right\} \right], \quad (10.26)$$

and,

$$b = \frac{(d_1 w_5 + d_2 w_6 + d_3 w_7 + w_4)(X - Y)}{\omega_v + \xi_v + \mu}, \quad (10.27)$$

where

$$D_a = \Pi(\omega_v + \xi_v + \mu), C_a = 2\mu\beta_p(w_4 + d_1 w_5 + d_2 w_6 + d_3 w_7),$$

and

$$P = \mu(v_3 w_3 + v_3 w_5 + v_3 w_6 + v_3 w_7 + v_8 + v_9 + v_3 w_4 + v_3 w_8 \varepsilon_n + v_3 w_9 \varepsilon_{nv}),$$

$$Q = \xi_v(v_3 w_3 + v_3 w_5 + v_3 w_6 + v_3 w_7 + v_8 w_8 + v_9 w_9 + v_1 w_1 + v_3 w_4$$

$$+ v_3 w_2 \mu + v_3 w_8 \varepsilon_n + v_3 w_9 \varepsilon_{nv} + v_2 w_1 \varepsilon_v + v_2 w_3 \varepsilon_v + v_2 w_4 \varepsilon_v + v_2 w_5 \varepsilon_v$$

$$+ v_2 w_6 \varepsilon_v + v_2 w_7 \varepsilon_v + v_2 w_8 \varepsilon_v + v_2 w_9 \varepsilon_v),$$

$$R = \omega_v(v_2 w_2 + v_3 w_3 + v_3 w_5 + v_3 w_6 + v_3 w_7 + v_8 w_8 + v_9 w_9 + v_3 w_4 + v_3 w_2 \varepsilon_v \\ + v_3 w_8 \varepsilon_n + v_3 w_9 \varepsilon_{nv}),$$

$$S = \mu(w_3 v_1 + w_5 v_1 + w_6 v_1 + w_7 v_1 + w_8 v_1 + w_9 v_1 + v_1 w_4 + v_8 w_8 \varepsilon_n \\ + v_9 w_9 \varepsilon_{nv} + v_2 w_2 \varepsilon_v), \quad (10.28)$$

$$T = \xi_v(v_2 w_1 + v_2 w_3 + v_2 w_4 + v_2 w_5 + v_2 w_6 + v_2 w_7 + v_2 w_8 + v_2 w_9 \\ + v_3 w_1 \varepsilon_v + v_3 w_3 \varepsilon_v + v_3 w_4 \varepsilon_v + v_3 w_5 \varepsilon_v + v_3 w_6 \varepsilon_v + v_3 w_7 \varepsilon_v + v_3 w_8 \varepsilon_v \\ + v_3 w_9 \varepsilon_v + v_8 w_8 \varepsilon_n + v_9 w_9 \varepsilon_{nv}),$$

$$U = \omega_v(v_1 w_2 + v_1 w_3 + v_1 w_4 + v_1 w_5 + v_1 w_6 + v_1 w_7 + v_1 w_8 + v_1 w_9 + v_2 w_2 \varepsilon_v \\ + v_8 w_8 \varepsilon_n + v_9 w_9 \varepsilon_{nv}),$$

$$X = \varepsilon_v v_2 \xi_v + \mu v_3 + w_v v_3 + v_3 \xi_v,$$

$$Y = \varepsilon_v v_3 \xi_v + \mu v_1 + w_v v_1 + v_2 \xi_v,$$

$$\beta_s = d_1 \beta_p, \beta_a = d_2 \beta_p, \beta_h = d_3 \beta_p.$$

It follows from Item (i) of Theorem 4.1 of [9] that the vaccination model (10.1) will undergo a backward bifurcation at $\mathbb{R}_{cv} = 1$ whenever the bifurcation coefficients, a and b (given by (10.26) and (10.27), respectively), are positive. It can be shown that the bifurcation coefficient b is automatically positive as follows. First of all, using the definitions for X and Y in (10.28), the quantity $X - Y$ can be simplified to (by using $v_1 = v_3 = 1$ as mentioned above):

$$X - Y = \xi_v(1 - \varepsilon_v)(1 - v_2), \quad (10.29)$$

which is positive since $\xi_v > 0$, $0 < \varepsilon_v < 1$ and the eigenvector $0 < v_2 < 1$ (from (10.22)). Thus, since $X - Y > 0$, $d_1 > 0$, $d_2 > 0$, $d_3 > 0$ and the eigenvectors w_4 , w_5 , w_6 , and w_7 are positive, it follows from (10.27) that the bifurcation coefficient b is automatically positive. Hence, since the bifurcation coefficient b is always positive, we only need to show that the coefficient a is positive for

backward bifurcation to occur. In particular, it can be shown from Eq. (10.26), and noting the definitions in (10.28) and the expressions for the eigenvectors in (10.22) and (10.23), that the backward bifurcation coefficient a is positive provided the following inequality holds:

$$P + Q + R > S + T + U. \quad (10.30)$$

Thus, it follows from Item (i) of Theorem 4.1 of [9]) that the vaccination model (10.1) will undergo a backward bifurcation at $\tilde{\mathbb{R}}_{cv} = 1$ whenever inequality (10.30) holds. \square

Appendix 2: Proof of Theorem 4

Proof Consider the vaccination model (10.1) with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ and $\tilde{\mathbb{R}}_{cv} \leq (1 - f_v) < 1$ (with f_v as defined in Eq. (10.6)). Furthermore, consider the following linear Lyapunov function:

$$\mathcal{L} = E + a_1 I_p + a_2 I_s + a_3 I_a + a_4 I_h,$$

where

$$a_1 = \frac{1}{\sigma_p + \delta_p + \mu} [\beta_p + a_2 r \sigma_p + a_3 (1 - r) \sigma_p], \quad a_2 = \frac{(a_4 \phi_s + \beta_s)}{\phi_s + \gamma_s + \delta_s + \mu},$$

$$a_3 = \frac{\beta_a}{\gamma_a + \delta_a + \mu}, \quad \text{and} \quad a_4 = \frac{\beta_h}{\gamma_h + \delta_h + \mu}.$$

It follows that the Lyapunov derivative is given by:

$$\dot{\mathcal{L}} = \dot{E} + a_1 \dot{I}_p + a_2 \dot{I}_s + a_3 \dot{I}_a + a_4 \dot{I}_h,$$

so that, upon substituting the equations for the respective derivatives of the model (10.1):

$$\begin{aligned} \dot{\mathcal{L}} = & \left[\beta_p \frac{S}{N} - a_1 (\sigma_p + \delta_p + \mu) + a_2 r \sigma_2 + a_3 (1 - r) \sigma_2 \right] I_p \\ & + \left[\beta_s \frac{S}{N} - a_2 (\phi_s + \gamma_s + \delta_s + \mu) + a_4 \phi_s \right] I_s + \left[\beta_a \frac{S}{N} - a_3 (\gamma_a + \delta_a + \mu) \right] I_a \\ & + \left[\beta_h \frac{S}{N} - a_4 (\gamma_h + \delta_h + \mu) \right] I_h + (\sigma_E + \mu) \left(\frac{\tilde{\mathbb{R}}_{cv}}{1 - f_v} - 1 \right) E, \end{aligned}$$

from which it follows that (noting that $S(t) \leq N(t)$ for all t in Ω),

$$\dot{\mathcal{L}} \leq (\sigma_E + \mu) \left(\frac{\hat{\mathbb{R}}_{cv}}{1 - f_v} - 1 \right) E.$$

Hence, $\dot{\mathcal{L}} \leq 0$ if $\hat{\mathbb{R}}_{cv} \leq 1 - f_v < 1$, and $\dot{\mathcal{L}} = 0$ if and only if $E(t) = 0$. Substituting $E(t) = 0$ into the equations of the vaccination model (10.1) shows that $(S(t), V(t), E(t), I_p(t), I_s(t), I_a(t), I_h(t), R_n(t), R_{nv}(t)) \rightarrow (S^*, V^*, 0, 0, 0, 0, 0, 0, 0)$ as $t \rightarrow \infty$. Furthermore, it can be shown that the largest compact invariant set in $\{(S(t), V(t), E(t), I_p(t), I_s(t), I_a(t), I_h(t), R_n(t), R_{nv}(t)) \in \Omega : \dot{\mathcal{L}} = 0\}$ is the disease-free equilibrium (\mathcal{E}_0) . Hence, it follows, by LaSalle's invariance principle [34], that the disease-free equilibrium (\mathcal{E}_0) of the vaccination model (10.1) is globally asymptotically stable in Ω whenever $\hat{\mathbb{R}}_{cv} < 1$. \square

Appendix 3: Proof of Theorem 5

Before proving this result, it is necessary to establish the following intermediate results.

*Proof of Positive Invariance and Attractivity of Ω_{**}*

Since $N(t) \leq \Pi/\mu$ for all t in Ω_{**} , it follows from the first equation of the vaccination model (10.1) that:

$$\begin{aligned} \frac{dS}{dt} &\leq \Pi + \omega_v V - (\xi_v + \mu)S, \\ &\leq \Pi + \left(\frac{\Pi}{\mu} - S \right) \omega_v - (\xi_v + \mu)S, \\ &\leq \frac{\Pi}{\mu} (\mu + \omega_v) - (\mu + \xi_v + \omega_v) S, \\ &\leq (\mu + \xi_v + \omega_v) (S^* - S). \end{aligned}$$

Hence, if $S(t) > S^*$, then $\frac{dS}{dt}$ is negative. Thus, $S(t) \leq S^*$ for all t , provided that $S(0) \leq S^*$. Using similar approach for the second equation of the vaccination model (10.1), and using the above bound, leads to the following bound:

$$\frac{dV}{dt} \leq \xi_v S^* - (\omega_v + \mu)V,$$

$$\begin{aligned} &\leq \xi_v \left[\frac{\Pi(\mu + \omega_v)}{\mu(\mu + \xi_v + \omega_v)} \right] - (\omega_v + \mu)V, \\ &\leq (\omega_v + \mu)(V^* - V). \end{aligned}$$

Following the same argument as above, we have $V(t) \leq V^*$ for all t , provided that $V(0) \leq V^*$. It follows from these bounds that:

$$\Omega_{**} = \{(S, V, E, I_p, I_s, I_a, I_h, R_n, R_{nv}) \in \Omega : S \leq S^*, V \leq V^*\} \quad (10.31)$$

is positively invariant and attracts all initial solutions in Ω_{**} .

Next-Generation Matrices for the Second Special Case of the Model

For the aforementioned (second) special case of the model (10.1), the associated next-generation matrix of new infection terms, denoted by F , is as given in Eq.(10.5), and the associated next-generation matrix of linear transition terms, denoted by \hat{V} , is given by:

$$\hat{V} = \begin{bmatrix} \hat{K}_1 & 0 & 0 & 0 & 0 \\ -\sigma_E & \hat{K}_2 & 0 & 0 & 0 \\ 0 & -r\sigma_p & \hat{K}_3 & 0 & 0 \\ 0 & -(1-r)\sigma_p & 0 & \hat{K}_4 & 0 \\ 0 & 0 & -\phi_s & 0 & \hat{K}_5 \end{bmatrix}, \quad (10.32)$$

with

$$\hat{K}_1 = \sigma_E + \mu, \hat{K}_2 = \sigma_p + \mu, \hat{K}_3 = \phi_s + \gamma_s + \mu, \hat{K}_4 = \gamma_a + \mu \text{ and } \hat{K}_5 = \gamma_h + \mu.$$

Proof of Theorem 5

Proof Consider the vaccination model (10.1) with $\delta_p = \delta_s = \delta_a = \delta_h = 0$ and $\varepsilon_n = \varepsilon_{nv} = 1$. We further assume that $\hat{\mathbb{R}}_{cv} < 1$. The proof is also based on using a comparison theorem [31, 46, 66]. Here, too, the equations for the infected compartments of the special case of the model (10.1) can be re-written as:

$$\frac{d}{dt} \begin{bmatrix} E(t) \\ I_p(t) \\ I_s(t) \\ I_a(t) \\ I_h(t) \end{bmatrix} = (F - \hat{V}) \begin{bmatrix} E(t) \\ I_p(t) \\ I_s(t) \\ I_a(t) \\ I_h(t) \end{bmatrix} - \hat{M} \begin{bmatrix} E(t) \\ I_p(t) \\ I_s(t) \\ I_a(t) \\ I_h(t) \end{bmatrix}, \quad (10.33)$$

where the matrices F and \hat{V} are as given in Eqs. (10.5) and (10.32), respectively, and the matrix \hat{M} (with S^* and V^* are as defined in Sect. 10.3.1) is given by:

$$\hat{M} = [(S^* - S) + (1 - \varepsilon_v)(V^* - V)] \begin{bmatrix} 0 & \beta_p & \beta_s & \beta_a & \beta_h \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \quad (10.34)$$

Since $S \leq S^*$, $V \leq V^*$ for all $t > 0$ in Ω_{**} (as shown above), it follows that the matrix \hat{M} , defined in Eq. (10.34), is non-negative. Hence, Eq. (10.33) can be re-written in terms of the following inequality:

$$\frac{d}{dt} \begin{bmatrix} E(t) \\ I_p(t) \\ I_s(t) \\ I_a(t) \\ I_h(t) \end{bmatrix} \leq (F - \hat{V}) \begin{bmatrix} E(t) \\ I_p(t) \\ I_s(t) \\ I_a(t) \\ I_h(t) \end{bmatrix}. \quad (10.35)$$

It should be recalled from the local asymptotic stable result for the disease-free equilibrium of the vaccination model (10.1) (given in Theorem 2) that all eigenvalues of the associated next-generation matrix FV^{-1} are negative if $\mathbb{R}_{cv} < 1$ (i.e., $F - V$ is a stable matrix). It follows that the eigenvalues of the next-generation matrix $F\hat{V}^{-1}$, associated with this special case of the model (10.1), are also negative if $\hat{\mathbb{R}}_{cv} < 1$ (i.e., $F - \hat{V}$ is also a stable matrix). Thus, the linearized differential inequality system (10.35) is stable whenever $\rho(F\hat{V}^{-1}) < 1$. Consequently [32, 61, 66, 71],

$$(E(t), I_p(t), I_s(t), I_a(t), I_h(t)) \rightarrow (0, 0, 0, 0, 0), \text{ as } t \rightarrow \infty.$$

Substituting $E(t) = I_p(t) = I_s(t) = I_a(t) = I_h(t) = 0$ into the differential equations for the rate of change of the $R_n(t)$, $R_{nv}(t)$, $V(t)$, and $S(t)$ compartments of the model (10.1) shows that:

$$R_n(t) \rightarrow 0, R_{nv}(t) \rightarrow 0, V(t) \rightarrow V^* \text{ and } S(t) \rightarrow S^*, \text{ as } t \rightarrow \infty.$$

Thus, the DFE (\mathcal{E}_0) of the second special case of the model (10.1) (with $\delta_p = \delta_s = \delta_a = \delta_h = 0$ and $\varepsilon_n = \varepsilon_{nv} = 1$) is globally asymptotically stable in Ω_{**} whenever $\tilde{\mathbb{R}}_{cv} < 1$. \square

Appendix 4: Proof of Theorem 7

Proof Consider the special case of the model (10.1) with $\varepsilon_v = \varepsilon_n = \varepsilon_{nv} = 1$ and $\omega_v = \delta_p = \delta_s = \delta_a = \delta_h = 0$. Setting $\delta_p = \delta_s = \delta_a = \delta_h = 0$ into the model shows that $N(t) \rightarrow N^* = \Pi/\mu$ as $t \rightarrow \infty$. For the rest of the analysis in this appendix, $N(t)$ will be replaced by the limiting value, N^* . It should be recalled that it has been shown (in Theorem 6) that this special case of the model has a unique endemic equilibrium (denoted by $\tilde{\mathcal{E}}_1$) whenever the associated reproduction number, denoted by $\tilde{\mathbb{R}}_v$, exceeds one. The proof of Theorem 7 will now be based on using a Krasnoselskii sub-linearity trick introduced by Hethcote and Thieme [37] (see also [21, 22, 55, 67, 76]). First of all, since $N(t) = N^*$, the following relation holds:

$$S(t) = N^* - [V(t) + E(t) + I_p(t) + I_s(t) + I_a(t) + I_h(t) + R_n(t) + R_{nv}(t)] \quad (10.36)$$

Substituting (10.36) into the model (10.1) gives the following reduced model:

$$\begin{cases} \dot{V} = \xi_v(N^* - V - E - I_p - I_s - I_a - I_h - R_n - R_{nv}) - (\omega_v + \mu)V, \\ \dot{E} = \tilde{\lambda}(N^* - V - E - I_p - I_s - I_a - I_h - R_n - R_{nv}) - (\sigma_E + \mu)E, \\ \dot{I}_p = \sigma_E E - (\sigma_p + \mu)I_p, \\ \dot{I}_s = r\sigma_p I_p - (\phi_s + \gamma_s + \mu)I_s, \\ \dot{I}_a = (1 - r)\sigma_p I_p - (\gamma_a + \mu)I_a, \\ \dot{I}_h = \phi_s I_s - (\gamma_h + \mu)I_h, \\ \dot{R}_n = \gamma_s I_s + \gamma_a I_a + \gamma_h I_h - (\xi_v + \mu)R_n, \\ \dot{R}_{nv} = \xi_v R_n - \mu R_{nv}. \end{cases} \quad (10.37)$$

where the associated *force of infection*, $\tilde{\lambda}$, is given by:

$$\tilde{\lambda} = \left(\frac{\beta_p I_p + \beta_s I_s + \beta_a I_a + \beta_h I_h}{N^*} \right). \quad (10.38)$$

The unique endemic equilibrium associated with the reduced system (10.37) now has the form:

$$\tilde{\mathcal{E}}_1 = (V^{**}, E^{**}, I_p^{**}, I_s^{**}, I_a^{**}, I_h^{**}, R_n^{**}, R_{nv}^{**}). \quad (10.39)$$

Linearizing the reduced model (10.37), around the endemic equilibrium ($\tilde{\mathcal{E}}_1$), gives the following linearized system:

$$\begin{cases} \dot{V} = -(\xi_v + c_2)V - \xi_v E - \xi_v I_p - \xi_v I_s - \xi_v I_a - \xi_v I_h - \xi_v R_n - \xi_v R_{nv}, \\ \dot{E} = -b_1 V - (b_1 + c_3)E + \sum_{j=\{p,s,a,h\}} (b_j - b_1)I_j - b_1 R_n - b_1 R_{nv}, \\ \dot{I}_p = \sigma_E E - c_4 I_p, \\ \dot{I}_s = r\sigma_p I_p - c_5 I_s, \\ \dot{I}_a = (1-r)\sigma_p I_p - c_6 I_a, \\ \dot{I}_h = \phi_s I_s - c_7 I_h, \\ \dot{R}_n = \gamma_s I_s + \gamma_a I_a + \gamma_h I_h - c_1 R_n, \\ \dot{R}_{nv} = \xi_v R_n - \mu R_{nv}. \end{cases} \quad (10.40)$$

where $b_1 = \frac{\beta_p I_p^{**} + \beta_s I_s^{**} + \beta_a I_a^{**} + \beta_h I_h^{**}}{N^*}$, $b_j = \frac{\beta_j S^{**}}{N^*}$ for $j = \{p, s, a, h\}$, and $c_i (i = 1, \dots, 7)$ are as defined in Sect. 10.3.2.1. The Jacobian associated with the linearized system (10.40) is given by:

$$J(\tilde{E}_1) = \begin{bmatrix} -(\xi_v + c_2) & -\xi_v \\ -b_1 & -(b_1 + c_3) & \alpha_p & \alpha_s & \alpha_a & \alpha_h & -b_1 & -b_1 \\ 0 & \sigma_E & -c_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & r\sigma_p & -c_5 & 0 & 0 & 0 & 0 \\ 0 & 0 & (1-r)\sigma_p & 0 & -c_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi_s & 0 & -c_7 & 0 & 0 \\ 0 & 0 & 0 & \gamma_s & \gamma_a & \gamma_h & -c_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \xi_v & -\mu \end{bmatrix}$$

with $\alpha_j = b_j - b_1$ (for $j = \{p, s, a, h\}$). Suppose, now, that the linearized system (10.40) has solution of the form [21, 22, 37, 55, 67, 76]:

$$\mathbf{Z}(t) = \mathbf{Z}_0 e^{\theta t}, \quad (10.41)$$

with $\mathbf{Z}_0 \in \mathbb{C} - \{0\}$, where $\mathbf{Z}_0 = (Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8)$, $\theta, Z_i \in \mathbb{C}$ (for $i = 1, 2, \dots, 8$) and \mathbb{C} denotes the complex numbers [21, 22, 37, 55, 67, 76]. Substituting a solution of the form (10.41) into the linearized system (10.40) gives:

$$\begin{aligned} \theta Z_1 &= -(c_1 + \xi_v)Z_1 - \xi_v(Z_2 + Z_3 + Z_4 + Z_5 + Z_6 + Z_7 + Z_8), \\ \theta Z_2 &= -b_1 Z_1 - (b_1 + c_3)Z_2 + \alpha_p Z_3 + \alpha_s Z_4 + \alpha_a Z_5 + \alpha_h Z_6 - b_1 Z_7 - b_1 Z_8, \\ \theta Z_3 &= \sigma_E Z_2 - c_4 Z_3, \\ \theta Z_4 &= r\sigma_p Z_3 - c_5 Z_4, \\ \theta Z_5 &= (1-r)\sigma_p Z_3 - c_6 Z_5, \\ \theta Z_6 &= \phi_s Z_4 - c_7 Z_6, \\ \theta Z_7 &= \gamma_s Z_4 + \gamma_a Z_5 + \gamma_h Z_6 - c_1 Z_7, \\ \theta Z_8 &= \xi_v Z_7 - \mu Z_8. \end{aligned} \quad (10.42)$$

System (10.42) can be simplified by moving all the negative terms in the last six equations of (10.42) to the respective left-hand sides [21, 22, 37, 55, 67, 76]. Further, the last six equations are then re-written in terms of Z_1 and substituted into the first two equations of (10.42), and all its negative terms are moved to the left-hand side as well. Finally, after adding the first and second equations of (10.42) and moving all the negative terms to the left-hand side, doing all these lead to the following system [21, 22, 37, 55, 67, 76]:

$$\begin{aligned} [1 + F_1(\theta)]Z_1 + [1 + F_2(\theta)]Z_2 &= (MZ)_1 + (MZ)_2, \\ [1 + F_3(\theta)]Z_3 &= (MZ)_3, \quad [1 + F_4(\theta)]Z_4 = (MZ)_4, \\ [1 + F_5(\theta)]Z_5 &= (MZ)_5, \quad [1 + F_6(\theta)]Z_6 = (MZ)_6, \\ [1 + F_7(\theta)]Z_7 &= (MZ)_7, \quad [1 + F_8(\theta)]Z_8 = (MZ)_8. \end{aligned} \tag{10.43}$$

where

$$\begin{aligned} F_1(\theta) &= \frac{\theta}{\xi_v + c_2} + \frac{a_1}{\xi_v + c_2}, \\ F_2(\theta) &= \frac{\theta}{\xi_v} + \frac{a_1 + c_3}{\xi_v} + \frac{\sigma_E}{(\theta + k_4)} \left(1 + \frac{a_1}{\xi_v}\right) + \frac{r\sigma_p\sigma_E}{(\theta + c_4)(\theta + c_5)} \left(1 + \frac{a_1}{\xi_v}\right) \\ &\quad + \frac{(1 - r)\sigma_p\sigma_E}{(\theta + c_4)(\theta + c_6)} \left(1 + \frac{a_1}{\xi_v}\right) + \frac{r\phi_s\sigma_p\sigma_E}{(\theta + c_4)(\theta + c_5)(\theta + c_7)} \left(1 + \frac{a_1}{\xi_v}\right) \\ &\quad + \frac{\sigma_E(\gamma_s + \gamma_a + \gamma_h)}{(\theta + c_1)(\theta + c_4)} \left(1 + \frac{a_1}{\xi_v}\right) + \frac{\sigma_E\xi_v(\gamma_s + \gamma_a + \gamma_h)}{(\theta + \mu)(\theta + c_1)(\theta + c_4)} \left(1 + \frac{a_1}{\xi_v}\right), \\ F_3(\theta) &= \frac{\theta}{c_4}, \quad F_4(\theta) = \frac{\theta}{c_5}, \quad F_5(\theta) = \frac{\theta}{c_6}, \quad F_6(\theta) = \frac{\theta}{c_7}, \quad F_7(\theta) = \frac{\theta}{c_1}, \quad F_8(\theta) = \frac{\theta}{\mu}, \end{aligned}$$

with

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\beta_p S^{**}}{k_3 N^*} & \frac{\beta_s S^{**}}{c_3 N^*} & \frac{\beta_a S^{**}}{c_3 N^*} & \frac{\beta_h S^{**}}{c_3 N^*} & 0 & 0 \\ 0 & \frac{\sigma_E}{c_4} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{r\sigma_p}{c_5} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{(1 - r)\sigma_p}{c_6} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\phi_s}{c_7} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\gamma_s}{c_1} & \frac{\gamma_a}{c_1} & \frac{\gamma_h}{c_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\xi_v}{\mu} & 0 \end{bmatrix}.$$

It can be verified that the endemic equilibrium $\tilde{E}_1 = (V^{**}, E^{**}, I_p^{**}, I_s^{**}, I_a^{**}, I_h^{**}, R_n^{**}, R_{nv}^{**})$ satisfies $\tilde{E}_1 = M\tilde{E}_1$ [21, 22, 37, 55, 67, 76]. The notation $(M\mathbf{Z})_i$ ($i = 1, 2, \dots, 8$) denotes the i th coordinate of the vector $M\mathbf{Z}$, and the matrix M has non-negative entries. If \mathbf{Z} is a solution of (10.43), then it is possible to find a minimal positive real number r such that [21, 22, 37, 55, 67, 76]

$$\|\mathbf{Z}\| = r\tilde{E}_1 \quad (10.44)$$

where $\|\mathbf{Z}\| = (\|Z_1\|, \|Z_2\|, \|Z_3\|, \|Z_4\|, \|Z_5\|, \|Z_6\|, \|Z_7\|, \|Z_8\|)$ with lexicographic order, and $\|\cdot\|$ is a norm in \mathbb{C} . The main goal is to show that $\text{Re}(\theta) < 0$. This is illustrated *via* the method of contradiction, as follows. Suppose, now, that $\text{Re}(\theta) \geq 0$ and consider the following two cases. \square

Case 1: $\theta = 0$

Setting $\theta = 0$ in (10.42) reduces it to a homogeneous linear system in the variables Z_i ($i = 1, \dots, 8$), with determinant given by:

$$\Delta = \left[\{\mu c_1 (c_2 + \xi_v) c_3 c_4 c_5 c_6 c_7\} \left(\frac{S^{**}(\tilde{\mathbb{R}}_v)}{N^*} - 1 \right) \right] - A_2 \quad (10.45)$$

where

$$\begin{aligned} A_2 = b_1 c_2 [\sigma_E (\mu + \xi_v) \{c_5 c_7 ((1-r)\gamma_a \sigma_p) + r c_6 \sigma_p (\gamma_h \phi_s + c_7 \gamma_s)\} \\ + \mu c_1 \{c_4 c_5 c_6 c_7 + \sigma_E (c_5 c_7 (c_6 + (1-r)\sigma_p) + r c_6 \sigma_p (c_7 + \phi_s))\}] > 0. \end{aligned}$$

To finally determine the sign of Δ , we need to determine the sign of $\left(\frac{S^{**}(\tilde{\mathbb{R}}_v)}{N^*} - 1 \right)$. This is explored below. Solving Eq. (10.37) at the endemic equilibrium \tilde{E}_1 gives:

$$\frac{S^{**}}{N^*} = \frac{c_3 E^{**}}{\beta_p I_p^{**} + \beta_s I_s^{**} + \beta_s I_a^{**} + \beta_h I_h^{**}}. \quad (10.46)$$

Substituting the expressions for $E^{**}, I_p^{**}, I_s^{**}, I_a^{**}$, and I_h^{**} from Eq. (10.15) into Eq. (10.46), and simplifying, gives (where $\tilde{\mathbb{R}}_v$ is as defined in Eq. (10.14)):

$$\frac{S^{**}}{N^*} = \frac{1}{\tilde{\mathbb{R}}_v}, \quad (10.47)$$

so that $\frac{S^{**}}{N^*} - \frac{1}{\tilde{\mathbb{R}}_v} = 0$. Thus, Eq. (10.45) now becomes (noting that $A_2 > 0$):

$$\Delta = -A_2 < 0.$$

Since the determinant (Δ) is negative, it follows that the system (10.42) has a unique solution, given by $\mathbf{Z} = \mathbf{0}$ (which corresponds to the disease-free equilibrium, \mathcal{E}_0).

Case 2: $\theta \neq 0$

Since we already assumed that $Re(\theta) > 0$, the remaining task is to show that the system has no non-trivial solution when $Re(\theta) > 0$. Clearly, we have that $F_j(\theta) > 0$, for all $j = \{p, s, a, h\}$, which implies that $|F_j(\theta) + 1| > 1$. We then define $F(\theta) = \min(|F_j(\theta) + 1|)$, for $j = \{p, s, a, h\}$. Then, $1 < F(\theta)$ and hence $\frac{r}{F(\theta)} < r$. Since r is a minimal positive real number such that $\|\mathbf{Z}\| \leq r \tilde{E}_1$ [21, 55, 67], which then implies that:

$$\|\mathbf{Z}\| > \frac{r}{F(\theta)} \tilde{E}_1. \quad (10.48)$$

On the other hand, by taking the norm of both sides of the third equation in (10.42) [21, 55, 67], and noting that M is a non-negative matrix, gives:

$$F(\theta)\|Z_3\| \leq |1+F_3(\theta)|\|Z_3\| = \|(MZ)_3\| \leq M\|Z_3\| \leq rM(\tilde{E}_1)_3 = r(\tilde{E}_1)_3 = rI_p^{**} \quad (10.49)$$

It follows from (10.49) that $\|Z_3\| \leq \frac{r}{F(\theta)} I_p^{**}$, which contradicts (10.48), due to the fact that r is minimal. Hence, $Re(\theta) < 0$. Thus, all eigenvalues of the characteristic equation associated with the linearized system (10.40) will have negative real part, so that the unique endemic equilibrium, \tilde{E}_1 , of the system (10.37) is locally asymptotically stable whenever $\tilde{\mathbb{R}}_v > 1$, as required. \square

References

1. Anderson, R.M.: The concept of herd immunity and the design of community-based immunization programmes. *Vaccine* **10**(13), 928–935 (1992)
2. Anderson, R.M., May, R.M.: Vaccination and herd immunity to infectious diseases. *Nature* **318**(6044), 323–329 (1985)
3. Banks, H.T., Davidian, M., Samuels, J.R., Sutton, K.L.: An inverse problem statistical methodology summary. In: Mathematical and Statistical Estimation Approaches in Epidemiology, pp. 249–302. Springer, Berlin (2009)
4. Bar-On, Y.M., Goldberg, Y., Mandel, M., Bodenheimer, O., Freedman, L., Kalkstein, N., Mizrahi, B., Alroy-Preis, S., Ash, N., Milo, R., et al.: Protection of BNT162b2 vaccine booster against COVID-19 in Israel. *N. Eng. J. Med.* **385**(15), 1393–1400 (2021)

5. Blayneh, K.W., Gumel, A.B., Lenhart, S., Clayton, T.: Backward bifurcation and optimal control in transmission dynamics of West Nile virus. *Bull. Math. Biol.* **72**(4), 1006–1028 (2010)
6. Blower, S.M., Dowlatabadi, H.: Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example. *Int. Stat. Rev./Revue Internationale de Statistique* **62**(2), 229–243 (1994)
7. Brozak, S.J., Pant, B., Saifdar, S., Gumel, A.B.: Dynamics of COVID-19 pandemic in India and Pakistan: A metapopulation modelling approach. *Infect. Dis. Model.* **6**, 1173–1201 (2021)
8. Carr, J.: Applications of Centre Manifold Theory, vol. 35. Springer, Berlin (2012)
9. Castillo-Chavez, C., Song, B.: Dynamical Models of Tuberculosis and their Applications. *Math. Biosci. Eng.* **1**(2), 361 (2004)
10. Chemaitelly, H., Ayoub, H.H., AlMukdad, S., Coyle, P., Tang, P., Yassine, H.M., Al-Khatib, H.A., Smatti, M.K., Hasan, M.R., Al-Kanaani, Z., et al.: Duration of mRNA vaccine protection against SARS-CoV-2 Omicron BA. 1 and BA. 2 subvariants in Qatar. *Nat. Commun.* **13**(1), 1–12 (2022)
11. Chowell, G.: Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. *Infect. Dis. Model.* **2**(3), 379–398 (2017)
12. Cuevas, E.: An agent-based model to evaluate the COVID-19 transmission risks in facilities. *Comput. Biol. Med.* **121**, 103827 (2020)
13. Curley, B.: “How long does immunity from COVID-19 vaccination last?” Healthline (2022). <https://www.healthline.com/health-news/how-long-does-immunity-from-covid-19-vaccination-last> Accessed 22 Mar 2022
14. Dan, J.M., Mateus, J., Kato, Y., Hastie, K.M., Yu, E.D., Faliti, C.E., Grifoni, A., Ramirez, S.I., Haupt, S., Frazier, A., et al.: Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science* **371**(6529), eabf4063 (2021)
15. Diekmann, O., Heesterbeek, J.A.P., Metz, J.A.: On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**(4), 365–382 (1990)
16. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**(5), 533–534 (2020)
17. Dushoff, J., Huang, W., Castillo-Chavez, C.: Backwards bifurcations and catastrophe in simple models of fatal diseases. *J. Math. Biol.* **36**(3), 227–248 (1998)
18. Eikenberry, S.E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., Kostelich, E., Gumel, A.B.: To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infect. Dis. Model.* **5**, 293–308 (2020)
19. Elbasha, E.H., Gumel, A.B.: Theoretical assessment of public health impact of imperfect prophylactic HIV-1 vaccines with therapeutic benefits. *Bull. Math. Biol.* **68**(3), 577–614 (2006)
20. Elbasha, E.H., Podder, C.N., Gumel, A.B.: Analyzing the dynamics of an SIRS vaccination model with waning natural and vaccine-induced immunity. *Nonlinear Anal. Real World Appl.* **12**(5), 2692–2705 (2011)
21. Esteva, L., Gumel, A.B., De León, C.V.: Qualitative study of transmission dynamics of drug-resistant malaria. *Math. Comput. Model.* **50**(3–4), 611–630 (2009)
22. Esteva, L., Vargas, C.: Influence of vertical and mechanical transmission on the dynamics of dengue disease. *Math. Biosci.* **167**(1), 51–64 (2000)
23. Ferguson, N.M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., et al.: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College COVID-19 Response Team, London **16** (2020)
24. Firth, J.A., Hellewell, J., Klepac, P., Kissler, S.: Using a real-world network to model localized COVID-19 control strategies. *Nat. Med.* 1–22 (2020). <https://doi.org/10.1038/s41591-020-1036-8>

25. Food, U., Administration, D., et al.: Vaccines and related biological products advisory committee meeting. FDA briefing document Moderna COVID-19 vaccine. Retrieved on February 5, 2021 (2020)
26. Food and Drug Administration and others: FDA briefing document, Pfizer-BioNTech COVID-19 vaccine. In: Vaccines and Related Biological Products Advisory Committee Meeting (2020)
27. Garba, S.M., Gumel, A.B.: Effect of cross-immunity on the transmission dynamics of two strains of dengue. *Int. J. Comput. Math.* **87**(10), 2361–2384 (2010)
28. Garba, S.M., Gumel, A.B., Bakar, M.A.: Backward bifurcations in dengue transmission dynamics. *Math. Biosci.* **215**(1), 11–25 (2008)
29. Gumel, A.B.: Causes of backward bifurcations in some epidemiological models. *J. Math. Anal. Appl.* **395**(1), 355–365 (2012)
30. Gumel, A.B., Iboi, E.A., Ngonghala, C.N., Elbasha, E.H.: A primer on using mathematics to understand COVID-19 dynamics: modeling, analysis and simulations. *Infect. Dis. Model.* **6**, 148–168 (2021)
31. Gumel, A.B., Iboi, E.A., Ngonghala, C.N., Ngwa, G.A.: Toward achieving a vaccine-derived herd immunity threshold for COVID-19 in the US. *Front. Public Health* **9** (2021)
32. Gumel, A.B., McCluskey, C.C., van den Driessche, P.: Mathematical study of a staged-progression HIV model with imperfect vaccine. *Bull. Math. Biol.* **68**(8), 2105–2128 (2006)
33. Gumel, A.B., Song, B.: Existence of multiple-stable equilibria for a multi-drug-resistant model of *Mycobacterium Tuberculosis*. *Math. Biosci. Eng.* **5**(3), 437 (2008)
34. Hale, J.K.: Ordinary Differential Equations. (1969)
35. Haynes, B., Messonnier, N.E., Cetron, M.S.: First travel-related case of 2019 novel coronavirus detected in United States: press release, Tuesday, January 21, 2020 (2020)
36. Hethcote, H.W.: The Mathematics of Infectious Diseases. *SIAM Rev.* **42**(4), 599–653 (2000)
37. Hethcote, H.W., Thieme, H.R.: Stability of the endemic equilibrium in epidemic models with subpopulations. *Math. Biosci.* **75**(2), 205–227 (1985)
38. Iboi, E., Sharomi, O.O., Ngonghala, C., Gumel, A.B.: Mathematical modeling and analysis of COVID-19 pandemic in Nigeria (2020). MedRxiv
39. Iboi, E.A., Gumel, A.B.: Mathematical assessment of the role of Dengvaxia vaccine on the transmission dynamics of dengue serotypes. *Math. Biosci.* **304**, 25–47 (2018)
40. Iboi, E.A., Ngonghala, C.N., Gumel, A.B.: Will an imperfect vaccine curtail the COVID-19 pandemic in the US? *Infect. Dis. Model.* **5**, 510–524 (2020)
41. Iboi, E.A., Okuneye, K., Sharomi, O., Gumel, A.B.: Comments on “A Mathematical Study to Control Visceral Leishmaniasis: An application to South Sudan”. *Bull. Math. Biol.* **80**(4), 825–839 (2018)
42. IHME COVID-19 health service utilization forecasting team: forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months (2020). MedRxiv
43. King, A.A., Domenech de Cellès, M., Magpantay, F.M., Rohani, P.: Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. R. Soc. B: Biol. Sci.* **282**(1806), 20150347 (2015)
44. Kissler, S.M., Tedijanto, C., Goldstein, E., Grad, Y.H., Lipsitch, M.: Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **368**(6493), 860–868 (2020)
45. La Salle, J.P.: The Stability of Dynamical Systems. SIAM (1976)
46. Lakshmikantham, V., Leela, S.: Differential and Integral Inequalities: Theory and Applications: Volume I: Ordinary Differential Equations. Academic Press, New York (1969)
47. Lakshmikantham, V., Vatsala, A.: Theory of differential and integral inequalities with initial time difference and applications. In: Analytic and Geometric Inequalities and Applications, pp. 191–203. Springer, Berlin (1999)
48. Liang, S.T., Liang, L.T., Rosen, J.M.: COVID-19: A comparison to the 1918 influenza and how we can defeat it (2021)
49. Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.m., Yuan, B., Kinoshita, R., Nishiura, H.: Incubation period and other epidemiological characteristics

- of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *J. Clin. Med.* **9**(2), 538 (2020)
50. Mahase, E.: COVID-19: Moderna vaccine is nearly 95% effective, trial involving high risk and elderly people shows. *Br. Med. J.* **371** (2020)
51. Mancuso, M., Eikenberry, S.E., Gumel, A.B.: Will vaccine-derived protective immunity curtail COVID-19 variants in the US? *Infect. Dis. Model.* **6**, 1110–1134 (2021)
52. Marino, S., Hogue, I.B., Ray, C.J., Kirschner, D.E.: A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J. Theor. Biol.* **254**(1), 178–196 (2008)
53. Mayo Clinic: COVID-19: Who's at higher risk of serious symptoms? (2022). <https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/coronavirus-who-is-at-risk/art-20483301> Accessed 18 Dec 2022
54. McLeod, R.G., Brewster, J.F., Gumel, A.B., Slonowsky, D.A.: Sensitivity and uncertainty analyses for a SARS model with time-varying inputs and outputs. *Math. Biosci. Eng.* **3**(3), 527–544 (2006)
55. Melesse, D.Y., Gumel, A.B.: Global asymptotic properties of an SEIRS model with multiple infectious stages. *J. Math. Anal. Appl.* **366**(1), 202–217 (2010)
56. Moghadas, S.M., Fitzpatrick, M.C., Sah, P., Pandey, A., Shoukat, A., Singer, B.H., Galvani, A.P.: The implications of silent transmission for the control of COVID-19 outbreaks. *Proc. Natl. Acad. Sci.* (2020). <https://doi.org/10.1073/pnas.2008373117>
57. Ngonghala, C.N., Goel, P., Kutor, D., Bhattacharyya, S.: Human choice to self-isolate in the face of the COVID-19 pandemic: a game dynamic modelling approach. *J. Theor. Biol.* **521**, 110692 (2021)
58. Ngonghala, C.N., Iboi, E., Eikenberry, S., Scotch, M., MacIntyre, C.R., Bonds, M.H., Gumel, A.B.: Mathematical assessment of the impact of non-pharmaceutical interventions on curtailing the 2019 novel coronavirus. *Math. Biosci.* **325**, 108364 (2020)
59. Ngonghala, C.N., Iboi, E.A., Gumel, A.B.: Could masks curtail the post-lockdown resurgence of COVID-19 in the US? *Math. Biosci.* **329**, 108452 (2020)
60. Ngonghala, C.N., Knitter, J.R., Marinacci, L., Bonds, M.H., Gumel, A.B.: Assessing the impact of widespread respirator use in curtailing COVID-19 transmission in the USA. *R. Soc. Open Sci.* **8**(9), 210699 (2021)
61. Ngonghala, C.N., Taboe, H.B., Safdar, S., Gumel, A.B.: Unraveling the dynamics of the Omicron and Delta variants of the 2019 coronavirus in the presence of vaccination, mask usage, and antiviral treatment. *Appl. Math. Model.* **114**, 447–465 (2023)
62. Oliver, S.E., Gargano, J.W., Marin, M., Wallace, M., Curran, K.G., Chamberland, M., McClung, N., Campos-Outcalt, D., Morgan, R.L., Mbaeyi, S., et al.: The Advisory Committee on Immunization Practices' interim recommendation for use of Pfizer-BioNTech COVID-19 vaccine—United States, December 2020. *Morb. Mortality Weekly Rep.* **69**(50), 1922 (2020)
63. Pearson, S.: What is the difference between the Pfizer, Moderna, and Johnson & Johnson COVID-19 vaccines? GoodRx. <https://www.goodrx.com/blog/comparing-COVID-19-vaccines/>. Accessed 25 June 2021
64. Pfizer: Pfizer and BioNTech to submit emergency use authorization request today to the US FDA for COVID-19 vaccine (2020)
65. Polack, F.P., Thomas, S.J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J.L., Marc, G.P., Moreira, E.D., Zerbini, C., et al.: Safety and efficacy of the BNT162b2 mRNA COVID-19 vaccine. *N. Eng. J. Med.* (2020)
66. Safdar, S., Ngonghala, C.N., Gumel, A.B.: Mathematical assessment of the role of waning and boosting immunity against the BA.1 Omicron variant in the United States. *Math. Biosci. Eng.* **20**(1), 179–212 (2023)
67. Safi, M.A., Gumel, A.B.: Global asymptotic dynamics of a model for quarantine and isolation. *Discrete Contin. Dyn. Syst. B* **14**(1), 209 (2010)
68. Sargent, J., Kumar, S., Buckley, K., McIntyre, J.: Johnson & Johnson announces real-world evidence and phase 3 data confirming substantial protection of single-shot COVID-19 vaccine in the US additional data show a booster increases protection1 (2021)

69. Schneider, K.A., Ngwa, G.A., Schwehm, M., Eichner, L., Eichner, M.: The COVID-19 pandemic preparedness simulation tool: Covidsim. *BMC Infect. Dis.* **20**(1), 1–11 (2020)
70. Self, W.H., Tenforde, M.W., Rhoads, J.P., Gaglani, M., Ginde, A.A., Douin, D.J., Olson, S.M., Talbot, H.K., Casey, J.D., Mohr, N.M., et al.: Comparative effectiveness of Moderna, Pfizer-BioNTech, and Janssen (Johnson & Johnson) vaccines in preventing COVID-19 hospitalizations among adults without immunocompromising conditions—United States, March–August 2021. *Morbidity Mortality Weekly Rep.* **70**(38), 1337 (2021)
71. Sharomi, O., Gumel, A.B.: Mathematical study of a risk-structured two-group model for Chlamydia transmission dynamics. *Appl. Math. Model.* **35**(8), 3653–3673 (2011)
72. Sidik, S.M.: Vaccines protect against infection from Omicron subvariant—but not for long. *Nature* (2022)
73. Srivastava, A., Chowell, G.: Understanding spatial heterogeneity of COVID-19 pandemic using shape analysis of growth rate curves (2020). *MedRxiv*
74. Tan, J.K., Leong, D., Munusamy, H., Zenol Ariffin, N.H., Kori, N., Hod, R., Periyasamy, P.: The prevalence and clinical significance of presymptomatic COVID-19 patients: how we can be one step ahead in mitigating a deadly pandemic. *BMC Infect. Dis.* **21**(1), 1–10 (2021)
75. Tariq, A., Lee, Y., Roosa, K., Blumberg, S., Yan, P., Ma, S., Chowell, G.: Real-time monitoring the transmission potential of COVID-19 in Singapore, March 2020. *BMC Med.* **18**, 1–14 (2020)
76. Thieme, H.R.: Local stability in epidemic models for heterogeneous populations. In: *Mathematics in Biology and Medicine*, pp. 185–189. Springer, Berlin (1985)
77. Thurner, S., Klimek, P., Hanel, R.: A network-based explanation of why most COVID-19 infection curves are linear. *Proc. Natl. Acad. Sci.* **117**(37), 22684–22689 (2020)
78. Tseng, H.F., Ackerson, B.K., Luo, Y., Sy, L.S., Talarico, C., Tian, Y., Bruxvoort, K., Tupert, J.E., Florea, A., Ku, J.H., et al.: Effectiveness of mRNA-1273 against SARS-CoV-2 Omicron and Delta variants (2022). *MedRxiv*
79. US Food and Drug Administration and others: FDA briefing document. In: *Oncology Drug Advisory Committee Meeting*, Silver Spring (2009)
80. US Food and Drug Administration and others: Coronavirus (COVID-19) update: FDA issues policies to guide medical product developers addressing virus variants. FDA. February **23**, 2021 (2021)
81. US Food and Drug Administration and others: FDA issues emergency use authorization for third COVID-19 vaccine. FDA, Silver Spring (2021)
82. Van den Driessche, P., Watmough, J.: Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **180**(1–2), 29–48 (2002)
83. Vermund, S.H., Pitzer, V.E.: Asymptomatic transmission and the infection fatality risk for COVID-19: implications for school reopening. *Clin. Infect. Dis.* **72**(9), 1493–1496 (2021)
84. Weintraub, K.: Enormous spread of Omicron' may bring 140M new COVID infections to US in the next two months, model predicts (2022). <https://www.wusa9.com/article/news/verify/how-long-does-it-take-for-the-vaccine-booster-to-get-to-full-protection/65-aa7344c2-fcd5-4c70-bbcd-046e9f697be7>. Accessed 22 Mar 2022
85. Wilder, B., Charpignon, M., Killian, J.A., Ou, H.C., Mate, A., Jabbari, S., Perrault, A., Desai, A., Tambe, M., Majumder, M.S., et al.: The role of age distribution and family structure on COVID-19 dynamics: a Preliminary Modeling Assessment for Hubei and Lombardy. Available at SSRN **3564800** (2020)
86. Worldometer.: COVID live update. Worldometer information. <https://www.worldometers.info/coronavirus/>. Accessed 2 April 2023
87. Xue, L., Jing, S., Miller, J.C., Sun, W., Li, H., Estrada-Franco, J.G., Hyman, J.M., Zhu, H.: A data-driven network model for the emerging COVID-19 epidemics in Wuhan, Toronto and Italy. *Math. Biosci.* **326**, 108391 (2020)

Chapter 11

Long-Term Dynamics of COVID-19 in a Multi-strain Model



Elisha B. Are, Jessica Stockdale, and Caroline Colijn

11.1 Introduction

COVID-19 persists as a major public health challenge in many countries. The emergence and spread of variants of concern (VOC) is an ongoing challenge to effective pandemic control. VOC have undermined control efforts through increased transmissibility [28, 35] and severity [38]. Omicron and its diversifying sub-lineages exhibit considerable and ongoing evasion of immunity against infection, though with an intrinsic severity that is lower than that of the Delta variant [22]. As it has become apparent that the disease will not be eradicated [1], continuous efforts towards improving and refining the understanding of the long-term dynamics of COVID-19 are important.

Mathematical modelling allows exploration of the long-term dynamics of COVID-19 given patterns in the emergence of VOC. A wide variety of modelling approaches have been used to explore COVID-19 dynamics and options for control [2, 21, 42]. Some mathematical modelling studies have specifically explored possible future longer-term dynamics of COVID-19. For instance [10] employed several models and predicted further emergence and spread of new variants in late 2021 to early 2022 and suggested that the impact of the emerging waves of infection can go unnoticed at the initial stage because of physical distancing measures. Similar studies have been conducted in different settings [3, 18, 40].

As SARS-CoV-2 continues to spread, new variants will keep emerging [26]. The continual evolution of the virus further complicates efforts to make long-term predictions for COVID-19 dynamics. For instance, pre-Omicron variants showed minimal ability to evade immunity, whereas Omicron and its sub-lineages emerged

E. B. Are (✉) · J. Stockdale · C. Colijn

Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada

e-mail: elisha_are@sfu.ca; jessica_stockdale@sfu.ca; ccolijn@sfu.ca

relatively suddenly and showed considerable ability to cause vaccine breakthrough infections and to escape infection-induced immunity [30, 36, 43].

Multi-strain modelling frameworks provide an appropriate tool for capturing complex disease dynamics when multiple strains are co-circulating or when resident strains are continuously replaced by more competent mutant strains. Multi-strain models have been used extensively to study the emergence and spread of new variants of SARS-CoV-2 [14, 17, 19]. In this chapter, we use a multi-strain model to analyse various long-term scenarios for the COVID-19 pandemic. We simulate the emergence of new variants using a Poisson process that determines the new variants' arrival times. We use a multivariate log-normal distribution to sample transmissibility and immune escape capacity of new variants, relative to the Omicron sub-lineage BA.2, informed by the evolution of variants to date. This modelling framework allows us to capture stochasticity in variants' arrival times and phenotypes and to explore the resulting long-term dynamics under multiple scenarios: lower escape capacity versus high transmission and vice versa, various arrival rates as well as the impact of reducing vaccine booster uptake in the future. We situate our work in the province of British Columbia, Canada. For model validation we present COVID-19 modelling projections that were done previously for the Omicron wave in six provinces in Canada, including British Columbia. The projections involve model calibration using a penalized maximum likelihood fit (described here) of model output to reported case data.

11.2 Methodology

We use a two-strain deterministic compartmental model that captures the dynamics of two SARS-CoV-2 variants. We develop a penalized maximum likelihood parameter estimation procedure to fit the model to past data and project COVID-19 cases in Canada during the Omicron wave, for model calibration (see Supplementary Information). Furthermore, we model long-term dynamics of COVID-19 in British Columbia (BC).

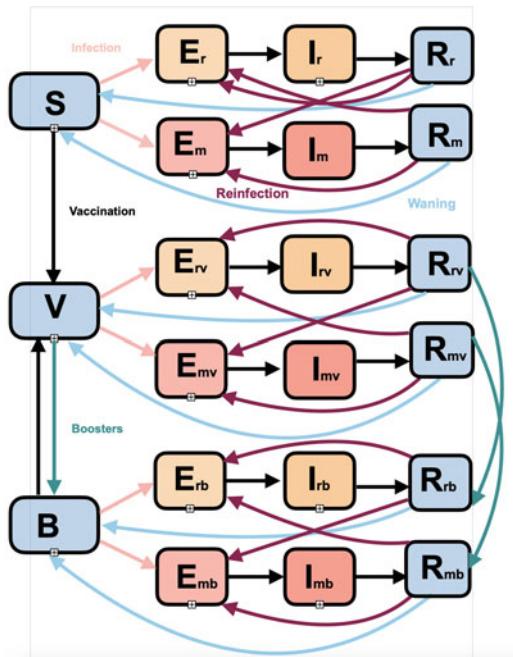
11.2.1 *Model Description*

In this model, the population is stratified into three streams—unvaccinated, vaccinated and boosted—those who have received a recent booster dose. It also allows the transmission of “resident” and “mutant” SARS-CoV-2 variants in the population (but does not account for the possibility of concurrent infection with both strains). In the rest of this section, for consistency, we use the term “strain” for SARS-CoV-2 variants. Each of the three population streams has seven compartments. The unvaccinated stream has the susceptibles (S), those exposed to the resident strain (E_r), those exposed to the mutant strain (E_m), those infected with the resident

strain (I_r), those infected with the mutant strain (I_m), those who recovered from resident strain (R_r) and those who recovered from the mutant strain (R_m). The vaccinated stream consists of the vaccinated (V), the vaccinated who are exposed to the resident strain (E_{rv}), vaccinated who are exposed to the mutant strain (E_{mv}), vaccinated individuals who are infected with the resident strain (I_{rv}), vaccinated individuals who are infected with the mutant strain (I_{mv}), vaccinated who recovered from the resident strain (R_{rv}) and those vaccinated who recovered from the mutant strain (R_{mv}). Similarly, the boosted stream has the boosted individuals (B), the boosted who are exposed to the resident strain (E_{rb}), boosted who are exposed to the mutant strain (E_{mb}), individuals who are boosted but are infected with the resident strain (I_{rb}), those who are boosted but are infected with the mutant strain (I_{mb}), boosted who recovered from the resident strain (R_{rb}) and boosted who recovered from the mutant strain (R_{mb}). The model assumes that individuals can acquire either the resident or the mutant strain according to the forces of infection λ_r and λ_m , defined below. Immunity wanes at the rate ω , with subscripts r , m and b , corresponding to the resident strain, mutant strain and the boosted, respectively. We assume different vaccine efficacy against infection for the vaccinated and boosted individuals. Furthermore, individuals are vaccinated or boosted at the rate v and b , respectively. The booster efficacy (against infection) is denoted by b_{eff} . Efficacy against infection is denoted by $1-\epsilon$, with subscripts corresponding to the respective strains. Moreover, those newly recovered have some protection against reinfection, which is denoted by $1-c_r$, $1-c_m$, $1-c_{rm}$ and $1-c_{mr}$, corresponding, respectively, to protection against infection by the resident strain after recovering from the resident strain, protection against infection by the mutant strain after recovering from the mutant strain, protection against infections by the resident strain after recovering from the mutant strain and protection against infections by the mutant strain after recovering from the resident strain. The rate σ at which individuals become infectious after being exposed is assumed to be the same regardless of the strain. To simplify the model further, we assume that the recovery rate γ does not depend on the strain. Moreover, we assume that the population size is constant over time by equating the birth rate to the background mortality rate (μ). The schematic diagram and the system of first-order ordinary differential equations governing the dynamics are presented in Fig. 11.1 and model (11.1).

$$\begin{aligned}\frac{dS}{dt} &= \mu N - (\lambda_r + \lambda_m)S - (v + \mu)S + (\omega_r R_r + \omega_m R_m) \\ \frac{dE_r}{dt} &= \lambda_r(S + c_r R_r + c_{rm} R_m) - (\sigma + \mu)E_r \\ \frac{dE_m}{dt} &= \lambda_m(S + c_{mr} R_r + c_m R_m) - (\sigma + \mu)E_m \\ \frac{dI_r}{dt} &= \sigma E_r - (\gamma + \mu)I_r\end{aligned}$$

Fig. 11.1 Model schematic. The compartments and flows illustrate the sub-populations and dynamics in the model



$$\frac{dI_m}{dt} = \sigma E_m - (\gamma + \mu) I_m$$

$$\frac{dR_r}{dt} = \gamma I_r - (c_r \lambda_r + c_{mr} \lambda_m + \mu + \omega_r) R_r$$

$$\frac{dR_m}{dt} = \gamma I_m - (c_{rm} \lambda_r + c_m \lambda_m + \mu + \omega_m) R_m$$

$$\frac{dV}{dt} = vS + \omega_b B + (\omega_r R_{rv} + \omega_m R_{mv}) - (bv_{be} + \mu + \epsilon_r \lambda_r + \epsilon_m \lambda_m) V$$

$$\frac{dE_{rv}}{dt} = \epsilon_r \lambda_r (V + c_r R_{rv} + c_{rm} R_{mv}) - (\sigma + \mu) E_{rv}$$

$$\frac{dE_{mv}}{dt} = \epsilon_m \lambda_m (V + c_{mr} R_{rv} + c_m R_{mv}) - (\sigma + \mu) E_{mv}$$

$$\frac{dI_{rv}}{dt} = \sigma E_{rv} - (\gamma + \mu) I_{rv}$$

$$\frac{dI_{mv}}{dt} = \sigma E_{mv} - (\gamma + \mu) I_{mv}$$

$$\frac{dR_{rv}}{dt} = \gamma I_{rv} - (c_r \epsilon_r \lambda_r + c_{mr} \epsilon_m \lambda_m + \mu + \omega_r + bv_{be}) R_{rv}$$

$$\begin{aligned}
\frac{dR_{mv}}{dt} &= \gamma I_{mv} - (c_{rm}\epsilon_r\lambda_r + c_m\epsilon_m\lambda_m + \mu + \omega_m + bv_{be})R_{mv} \\
\frac{dB}{dt} &= bv_{be}V + (\omega_b R_{rb} + \omega_m R_{mb}) - (1 - b_{eff})(\lambda_r + \lambda_m)B - (\mu + \omega_b)B \\
\frac{dE_{rb}}{dt} &= (1 - b_{eff})\lambda_r(B + c_r R_{rb} + c_{rm} R_{mb}) - (\sigma + \mu)E_{rb} \\
\frac{dE_{mb}}{dt} &= (1 - b_{eff})\lambda_m(B + c_{mr} R_{rb} + c_m R_{mb}) - (\sigma + \mu)E_{mb} \\
\frac{dI_{rb}}{dt} &= \sigma E_{rb} - (\gamma + \mu)I_{rb} \\
\frac{dI_{mb}}{dt} &= \sigma E_{mb} - (\gamma + \mu)I_{mb} \\
\frac{dR_{rb}}{dt} &= \gamma I_{rb} + bv_{be}R_{rv} - (1 - b_{eff})(c_r\lambda_r + c_{mr}\lambda_m)R_{rb} - (\mu + \omega_r)R_{rb} \\
\frac{dR_{mb}}{dt} &= \gamma I_{mb} + bv_{be}R_{mv} - (1 - b_{eff})(c_{rm}\lambda_r + c_m\lambda_m)R_{mb} - (\mu + \omega_m)R_{mb},
\end{aligned} \tag{11.1}$$

where $N = S + E_r + E_m + I_r + I_m + R_r + R_m + V + E_{rv} + E_{mv} + I_{rv} + I_{mv} + R_{rv} + R_{mv} + W + E_{rb} + E_{mb} + I_{rb} + I_{mb} + R_{rb} + R_{mb}$.

The forces of infection are $\lambda_r = \tau(t)\psi(t)\frac{\beta_r(I_r + I_{rv} + I_{rb})}{N}$ and $\lambda_m = \tau(t)\psi(t)\frac{\beta_m(I_m + I_{mv} + I_{mb})}{N}$. These are the standard forces of infection, modified to allow simulation of measures aiming to reduce transmission. For example, we model the implementation and relaxation of physical distancing measures with functions $\tau(t)$ and $\psi(t)$, respectively, where $\tau(t) = (1 - \tau_1/(1 + e^{(-s(t-d))}))$ and $\psi(t) = (1 + \psi_1/(1 + e^{(-s(t-d))}))$. The parameters $\tau_1, \psi_1 \in [0, 1]$ are, respectively, the level of reduction in transmission when the intervention is fully in place, and the increase in transmission when distancing measures are lifted. The parameter s is the steepness of the decline/increase in transmission, and d is the delay before relaxation/implementation of measures are 50% effective. This approach allows us to easily account for the impact of measures that reduce transmission (including masking, ventilation, distancing, work from home policies or other efforts). The choice of parameters is informed by historic data on physical distancing policy stringency [7].

11.2.2 Parameter Estimation

During the Omicron wave, PCR testing capacity was heavily impacted by rapidly rising infections. This led to significant changes in testing rate and reporting during

that period. Moreover, variant-specific data were sparse, making it challenging to identify the actual number of cases due to individual variants over time.

We develop a penalized maximum likelihood (or semi-Bayes) procedure [8] to fit the model to reported incidence data, in order to estimate some of the parameters of interest while taking these estimation challenges into account. We adopted penalized maximum likelihood methods because the classical maximum likelihood procedure yielded a relatively flat likelihood. We assume that the reported case count C_t is over-dispersed [12] and therefore follows a negative binomial distribution:

$$C_t \sim NegBin(\delta(t), \phi).$$

We define a log-likelihood function as follows:

$$\ln(\mathcal{L}) = \sum_t \ln\{\mathcal{P}(C_t | \hat{\theta})\},$$

where $\delta(t) = qt_p H$, and q is the reporting probability, $H = \sigma(E_r + E_{rv} + E_{rb} + E_m + E_{mv} + E_{mb})$ is the model predicted daily number of true infections, $\hat{\theta}$ are the fitted parameters (any combination of model parameters) and ϕ is the dispersion parameter, and we defined an age-corrected testing probability (t_p), which sprung from the idea of using the testing rate in those 70 years and older to infer testing rate in younger cohorts. Because those aged 70 years and older were prioritized for testing, their testing rates were relatively consistent.

In addition to the likelihood components described above, we introduce several penalties to direct our maximum likelihood procedure, using known statistics from during the fitting period. These statistics are (i) on date d_{prop} , a proportion r_{prop} of cases were observed to be caused by the resident strain (from genomic surveillance), and (ii) during a period t_{rate} , the growth rate of the mutant strain was observed to be m_{rate} . In the likelihood, we calculate the squared difference between the known statistics r_{prop} and m_{rate} , and these quantities as predicted by our model. This is used to set penalties p_{prop} and p_{rate} , respectively, within the penalized likelihood function.

More generally, any penalty p_{stat} could be defined that compares some quantity or statistic X output by the model to its observed value X_{stat} .

$$p_{stat} = (X_{stat} - X)^2$$

Essentially, parameters providing model fits and therefore statistics X that do not agree with the observed X_{stat} (here, r_{prop} or m_{rate}) result in a negative contribution to the likelihood value. Parameters providing model fits that strongly agree with the observed X_{stat} (here, r_{prop} or m_{rate}) lead to a positive contribution to the likelihood.

This gives a penalized log-likelihood function, to be numerically maximized, of the form

$$\ln(\mathcal{L}) - \frac{p_w(p_{\text{prop}} + p_{\text{rate}})}{2}, \quad (11.2)$$

where $\ln(\mathcal{L})$ is the log-likelihood and $\frac{p_w}{2}$ is a weight attributed to each penalty. This weight determines how impactful the penalties are in the maximum likelihood procedure: a high weight can be considered similar to placing a strong prior on the observed statistics r_{prop} and m_{rate} being true, even though we are not working in a full Bayesian framework. This setup allows us to find various parameter estimate combinations that maximize the penalized likelihood function, using the function *optim* in R, version 4.1.2 (2021-11-01), in RStudio version 2021.09.2+382.

11.2.3 Data Sources

The incidence case data used in this model is publicly available on the official [COVID-19 dashboard of the British Columbia Centre for Disease Control \(BCCDC\)](#) [6]. All other resources and code developed for this model are freely available on the GitHub repository: github.com/ElishaBayode/Omicron_projection.

11.3 COVID-19 Long-Term Scenarios Modelling

We modelled long-term COVID-19 scenarios under the assumption that variants will continue to emerge and spread. Moreover, we suppose that their arrival times in BC follow a Poisson process and that the emergent variants are phenotypically distinct from one another, with changes in transmissibility and immune evasion. The emergence and spread of SARS-CoV-2 variants of concern within BC can be modelled using the Poisson process framework. The underlying assumptions consider the average rate of arrival to be a known variable over a specific time frame, while the actual timing of arrival is a stochastic event. To capture this process in our model, we employ a Poisson distribution with a mean of λ , which yields a random distribution of arrival times across the 3-year simulation period. The strain swap framework described below is used to introduce new variants into the model at the arrival times generated from the Poisson process.

Known SARS-CoV-2 variants emerged in specific countries/regions and spread subsequently to other places. Whether and how quickly a variant will spread across borders depends on numerous factors including its growth advantage over existing variants and how connected its place of origin is to other parts of the world. Here we assume that new SARS-CoV-2 variants arrive in BC following a Poisson process with known arrival rate λ (0.0083 per day, i.e. on average every 4 months) over a 3-year period.

We assume that each arriving variant has a distinct phenotype relative to Omicron subvariant BA.2, which caused the second Omicron wave in Canada. We

considered increase in transmissibility and improved ability to evade vaccine- and infection-induced immunity against infection. While severe disease is of course an important driver of the impact of COVID-19 infections, we focus on transmission and immunity against infection because these impact the SARS-CoV-2 population dynamics. We do not focus on severe disease and the evolution of intrinsic or realized severity. Since the advent of Omicron, immunity is frequently discussed in the context of protection against severe disease; however, the burden of severe disease can be estimated from the number of infections if the probability of severe disease given infection is known.

Increases in immune evasion and transmissibility (in the form of multiplicative factors) are sampled from multivariate log-normal distributions. The log-normal distribution is appropriate in this context because we are interested in non-negative increases in fitness. The means of the distributions are chosen under various scenarios with references to published work on transmissibility and immune escape capacity of emerging variants to date [27, 32].

Combining arrival times from the Poisson process and the sampled increases from the log-normal distribution, for each arrival time, we simulate the impact of multiple SARS-CoV-2 variants, over the next 3 years, with distinct transmissibility and immune evading capacity.

Upon arrival of a new variant, we pause the BC transmission model and initiate a “variant swap”, whereby the new “mutant” moves into the population, the previous “mutant” becomes the “resident” strain and the old resident is assumed to no longer be circulating at measurable levels and is therefore removed from the model. This requires two main actions: (i) updating the resident and mutant strain model parameters and (ii) initializing the new mutant strain with a proportion of the current infections. To illustrate:

Variant Swap Function

Suppose that

$$\Psi(t) = \{\epsilon_r, \epsilon_m, \beta_r, \beta_m, \omega_r, \omega_m, c_r, c_m, c_{rm}, c_{mr}, \theta\}(t)$$

is the set of parameters in model (11.1) at time t , with ϵ , β , ω and c parameters as defined above, and θ representing the remaining set of non-variant-specific model parameters. Suppose also that

$$\Phi(t) = \{S(t), V(t), B(t), \Phi_r(t), \Phi_m(t)\}$$

is the status of system (11.1) at time t , where

$$\Phi_r(t) = \{E_r(t), I_r(t), R_r(t), E_{rv}(t), I_{rv}(t), R_{rv}(t), E_{rb}(t), I_{rb}(t), R_{rb}(t)\}$$

defines the status of all compartments concerning the resident strain and equivalently for $\Phi_m(t)$. Finally, let x and y be the new variant sampled increases from the multivariate log-normal distribution, which are increases in transmissibility and immune escape capacity, respectively. Then, at a given variant arrival time $t = n$

from the Poisson process, the variant swap is initialized by modifying (i) the model parameters in $\Psi(n)$ and (ii) the status of the model system in $\Phi(n)$. First, we set $\Psi(n)$ equal to

$$\begin{aligned}\hat{\Psi} = \{\dot{\epsilon}_r = \epsilon_m, \dot{\epsilon}_m = y\epsilon_m, \dot{\beta}_r = \beta_m, \dot{\beta}_m = x\beta_m, \dot{\omega}_r = \omega_m, \dot{\omega}_m = \omega_m, \dot{c}_r = c_m, \\ \dot{c}_m = c_m, \dot{c}_{rm} = c_{rm}, \dot{c}_{mr} = c_{mr}, \dot{\theta} = \theta\},\end{aligned}$$

which relabels the old mutant strain as the new resident and introduces the new variant with parameters as simulated from the multivariate log-normal. It would also be possible to sample new ω and c parameters for the new mutant strain, but we hold these constant for simplicity.

Second, we update $\Phi(n)$ using the status of the system one time-step prior to variant arrival, at $t = n - 1$. Rather than introducing a single instance of the new mutant strain (which may take a long time to take off in the population), we assume that a proportion ρ_i of the infections at time $n - 1$ were in fact already the new mutant, circulating undetected.

To model increased variant immune evasion, we additionally assume that a proportion ρ_s of individuals previously recovered from the old resident are fully susceptible to reinfection. The remaining proportion $1 - \rho_s$ are considered recovered from the new resident (old mutant) strain. This results in the following expression for $\Phi(n)$:

$$\begin{aligned}\Phi(n) = \{S(n) = S(n - 1) + \rho_s R_r(n - 1), V(n) = V(n - 1) + \rho_s R_{rv}(n - 1), \\ B(n) = B(n - 1) + \rho_s R_{rb}(n - 1), \Phi_r(n), \Phi_m(n)\},\end{aligned}$$

where

$$\begin{aligned}\Phi_r(n) = \{E_r(n) = (1 - \rho_i)(E_r(n - 1) + E_m(n - 1)), \\ I_r(n) = (1 - \rho_i)(I_r(n - 1) + I_m(n - 1)), \\ R_r(n) = R_m(n - 1) + (1 - \rho_s)R_r(n - 1), \\ E_{rv}(n) = (1 - \rho_i)(E_{rv}(n - 1) + E_{mv}(n - 1)), \\ I_{rv}(n) = (1 - \rho_i)(I_{rv}(n - 1) + I_{mv}(n - 1)), \\ R_{rv}(n) = R_{mv}(n - 1) + (1 - \rho_s)R_{rv}(n - 1), \\ E_{rb}(n) = (1 - \rho_i)(E_{rb}(n - 1) + E_{mb}(n - 1)), \\ I_{rb}(n) = (1 - \rho_i)(I_{rb}(n - 1) + I_{mb}(n - 1)), \\ R_{rb}(n) = R_{mb}(n - 1) + (1 - \rho_s)R_{rb}(n - 1)\} \\ \Phi_m(n) = \{E_m(n) = \rho_i(E_r(n - 1) + E_m(n - 1)), \\ I_m(n) = \rho_i(I_r(n - 1) + I_m(n - 1)),\end{aligned}$$

$$\begin{aligned}
R_m(n) &= 0, \\
E_{mv}(n) &= \rho_i(E_{rv}(n-1) + E_{mv}(n-1)), \\
I_{mv}(n) &= \rho_i(I_{rv}(n-1) + I_{mv}(n-1)), \\
R_{mv}(n) &= 0, \\
E_{mb}(n) &= \rho_i(E_{rb}(n-1) + E_{mb}(n-1)), \\
I_{mb}(n) &= \rho_i(I_{rb}(n-1) + I_{mb}(n-1)), \\
R_{mb}(n) &= 0\}.
\end{aligned}$$

We explore various long-term scenarios under continuous boosting versus discontinuation of boosting, low increase in escape capacity and higher increase in transmissibility and vice versa.

For the low increase in transmissibility with high increase in escape scenario, we assumed that the mean of increase in transmissibility is lower than the mean of increase in immune escape capacity in the multivariate log-normal model, and we assumed the converse for the high increase in transmissibility with low increase in escape capacity scenario. In all the figures presented, symptomatic cases and symptomatic infections are used interchangeably, which are simply the daily incident infections multiplied by a constant $q = 0.56$ (i.e. daily reported symptomatic cases).

11.4 Results

Our model calibration yielded a reasonable fit to data. With hindsight, our model predictions of a low to moderate resurgence of cases when control measures are lifted were largely accurate for all of the provinces and in Canada at large (see Figs. 11.8, 11.9 and 11.10 in the supplement). Baseline parameters description, values, and sources are presented in Table 11.1.

Table 11.1 Baseline parameter description, values and sources

Parameters	Description	Values	Sources
μ	Natural mortality rate	1/(82 years)	Canada life expectancy [29]
γ	Recovery rate	1/(6 days)	[23]
σ	Rate of progression from E to I	1/(3 days)	[41]
q	Ascertainment rate	56%	Assumed higher than in [33]
ρ_s	Proportion susceptible to mutant strain when swap function is applied	0.5	Assumed
ρ_i	Proportion of mutant strain when swap function is applied	0.001	Assumed

We present results of multiple runs of the model and a representative singular run, together with the output of the sampling from the multivariate log-normal distribution. Multiple model runs saw emergence of many waves of resurgence of cases with varying peak height and width, depending on the time interval between arrival dates of variants, transmissibility and immune escape capacity. As the pandemic progresses into 2024 with frequent introduction of new variants, subsequent waves peak at a lower level of infection compared to earlier waves in late 2022 and early 2023, in simulations that have shown major waves earlier. Variants with increased immune escape without a commensurate increase in transmission may have a slower growth rate than when significant increase occurs in both traits (Fig. 11.2).

We simulate a scenario where the increase in transmissibility is higher than the increase in immune escape capacity. Results of multiple simulations show that the wave peaks tend to be higher when the increase in transmissibility is higher.

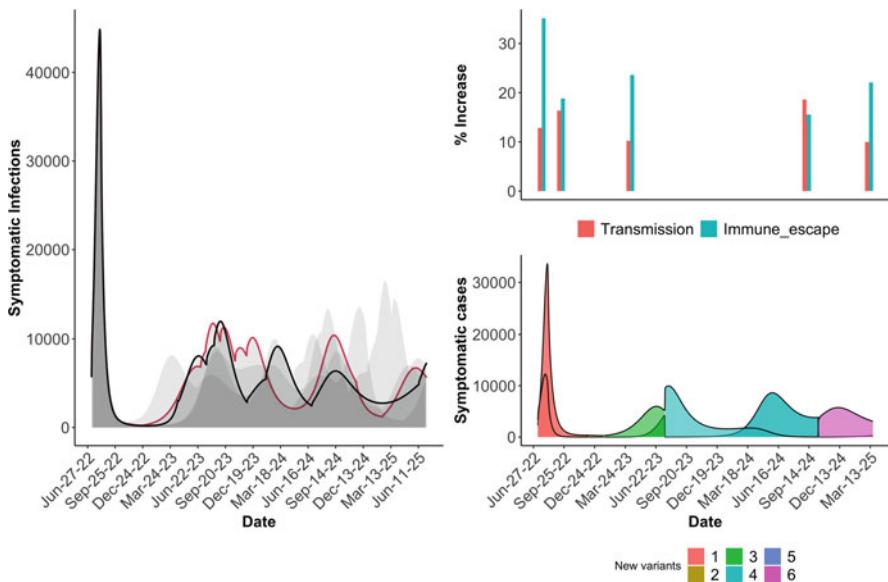


Fig. 11.2 Showing Scenario 1: lower transmissibility and higher immune escape capacity. The left plot shows multiple runs of the simulation (faded) with two illustrative trajectories (black and red), which are daily incident infections with constant reporting probability $q = 0.56$. The upper right plot shows the percentage increase in transmissibility (red) and percentage increase in immune escape capacity (blue), resulting from sampling the multivariate log-normal distribution. The lower right plot shows a representative simulation of mutant and resident variants over time (colour coded accordingly, by new mutant variant). These are also incident infections with constant reporting probability q . Baseline parameter values include $\mu = 1/82$ years, $\sigma = 1/3$ days, mean of increase in transmission $\beta_{inc} = 10\%$ and mean of increase in immune escape $\epsilon_{inc} = 22\%$, which are parameters of the multivariate normal distribution with variances 0.0041 and 0.0021, respectively. $c_m = 0.05$, $c_{mr} = 0.10$, $c_{rm} = 0.05$, $w_m = 0.0065$, $w_r = 0.0065$, $w_b = 0.0054$, $b = 0.0075$. See Table 11.1 for baseline parameter values

Moreover, a mutant variant with lower transmissibility can replace a resident with higher transmissibility if the latter is already on the decline (Fig. 11.3), assuming positive increase for all emergent variants. This is plausible, in practice, when an emerging variant can evade immunity acquired from previous infection by the resident.

Vaccines in the form of booster doses can be used to curb resurgence of COVID-19, both via protection against severe disease given infection and via protection against infection. However, as the pandemic continues, people's willingness to continue to get regular booster doses may dampen due to vaccination fatigue [39]. We analyse the impact of reduction in uptake of booster doses on the long-term dynamics of COVID-19. When the booster rate is reduced significantly, the peak heights are considerably higher (compare Figs. 11.3 and 11.4).

To further quantitatively explore the impact of viral evolution and boosting on COVID-19 dynamics in the next 3 years, we choose four summary statistics as metrics in subsequent analyses: the peak of infections, the interquartile range of

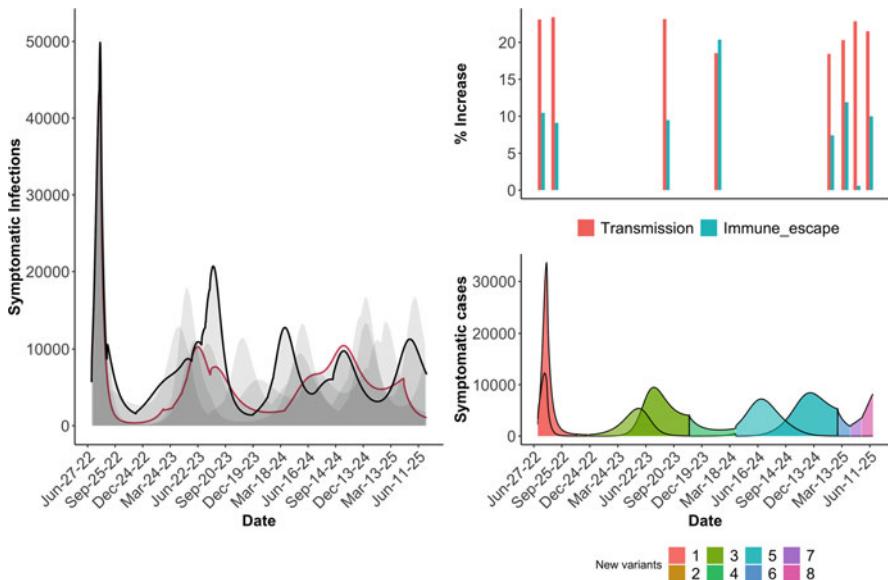


Fig. 11.3 Showing Scenario 2: higher transmissibility and lower immune escape capacity. The left plot shows results of multiple simulation runs (faded) with two illustrative trajectories (black and red). The plots show daily incident infections with a fixed reporting rate q . The upper right plot shows percentage increase in transmissibility (red) and percentage increase in immune escape capacity (blue), respectively, resulting from sampling the multivariate log-normal distribution. The lower right plot shows a representative simulation of mutant and residents variants over time (colour coded accordingly, by new mutant variant). Baseline parameter values include $\mu = 1/82$ years, $\sigma = 1/3$ days, mean of increase in transmission $\beta_{inc} = 22\%$ and mean of increase in immune escape $\epsilon_{inc} = 10\%$, which are parameters of the multivariate normal distribution with variances 0.0041 and 0.0021, respectively. $c_m = 0.05$, $c_{mr} = 0.10$, $c_{rm} = 0.05$, $w_m = 0.0065$, $w_r = 0.0065$, $w_b = 0.0054$, $b = 0.0075$ (and swap function parameters: $\rho_i = 0.001$, $\rho_s = 0.5$)

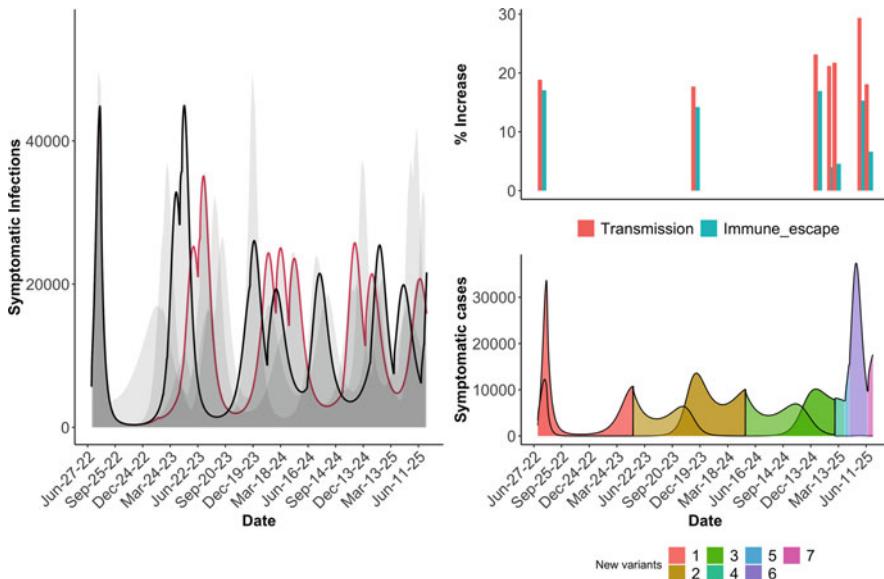


Fig. 11.4 Scenario 3: higher transmissibility and lower immune escape capacity, when the value of the boosting parameter b is reduced by 90%. Left: daily incident infections with a fixed reporting rate q . The left plot shows results of multiple runs of simulation (faded) with two illustrative trajectories (black and red). The upper right plot shows percentage increase in transmissibility (red) and percentage increase in immune escape capacity (blue), resulting from sampling the multivariate log-normal distribution. The lower right plot shows a representative simulation of mutant and resident variants over time (colour coded accordingly, by new mutant variant). Baseline parameter values are as presented in Fig. 11.3, save for b that is reduced by 90%

symptomatic infections, the total symptomatic infections and the variance in the total symptomatic infections. We employ violin plots for visualization (Figs. 11.5, 11.6 and 11.7). These analyses are done for projections starting from ending of October 2022. In these figures, note that the variance of the total asymptotic infection is measured in a different unit than that of the population size.

Figure 11.5 shows results of varying the Poisson process arrival rate from 2 to 6 months for 100 runs of the simulation. In Fig. 11.5a, the median global peak of symptomatic infection does not depend on arrival rate (Fig. 11.5a), though the distributions are very similar. A similar pattern is observed for the interquartile ranges, but the median of 2 and 4 months' arrival rates is slightly higher than for 6 months' arrival rate (Fig. 11.5b). In the same vein, total symptomatic infections are higher for 2 months arrival rate compared to 4 and 6 months' arrival rates (Fig. 11.5c). The variance in the total symptomatic infections is very similar across the three arrival rates.

In Fig. 11.6a, violin plots show the distribution of the highest peaks in 100 runs of the simulation, with embedded box plots showing the medians for two scenarios—when the mean and variance of increase in transmission is greater than those of the increase in immune escape capacity, as well as the reverse, for emerging variants.

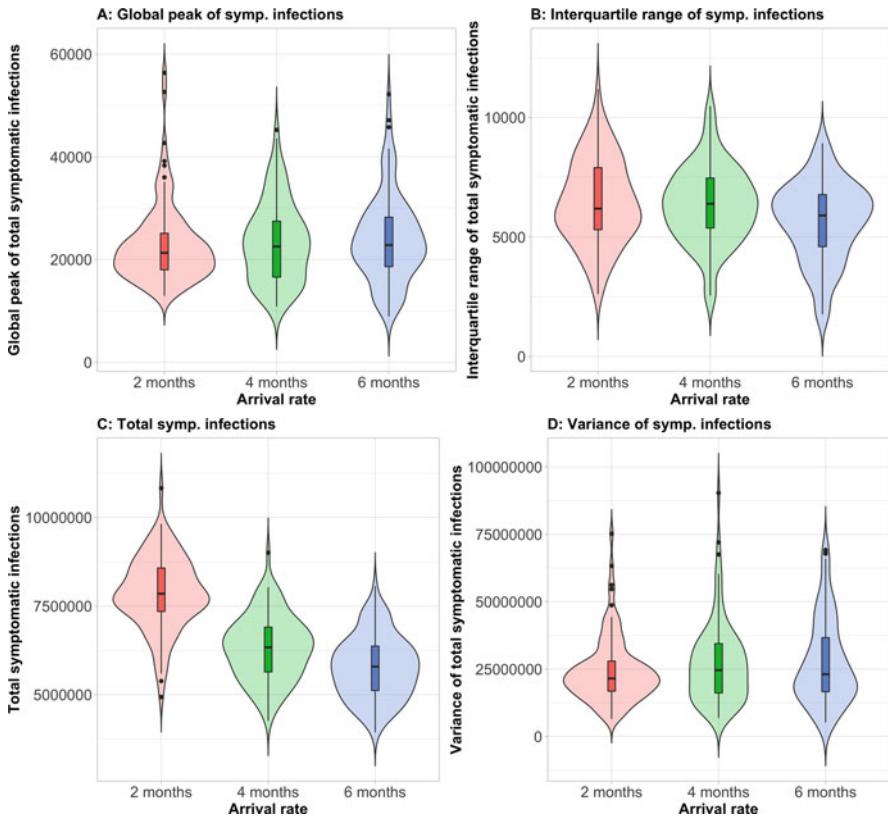


Fig. 11.5 Various summary statistics for different arrival rates of the Poisson process for variant arrival times: A. Violin plot of total symptomatic infections for arrival rates of 2, 4 or 6 months, with box plot embedded, for 100 runs of the simulation. B. Interquartile range of total symptomatic infections for multiple runs of the simulation and various arrival rates for the Poisson process. C. Total symptomatic infections with embedded box plots for different runs of the simulation for various arrival rates of the Poisson process. D. Variance of total symptomatic infections of multiple runs of the simulation. All model parameters are held constant except the arrival rate, which is varied. Here $\beta_{inc} = 35\%$, $\epsilon_{inc} = 10\%$ for the red violins, and the means and variances are reversed for the blue violins. Baseline parameters are kept as in Fig. 11.3

This suggests that when the increase in transmission is higher than the increase in immune escape capacity, the variant causes symptomatic infections with peaks that are frequently higher than when mutations favour higher immune escape capacity over transmissibility. Similarly, interquartile ranges are higher when the increase in transmission exceeds the increase in immune escape (Fig. 11.6b). This is also true for total symptomatic infections and variance of total symptomatic infections (see Fig. 11.6c and d).

Figure 11.7 shows the impact of possible discontinuation of boosting. For all four metrics that we analysed the pattern is consistent and intuitive—significant

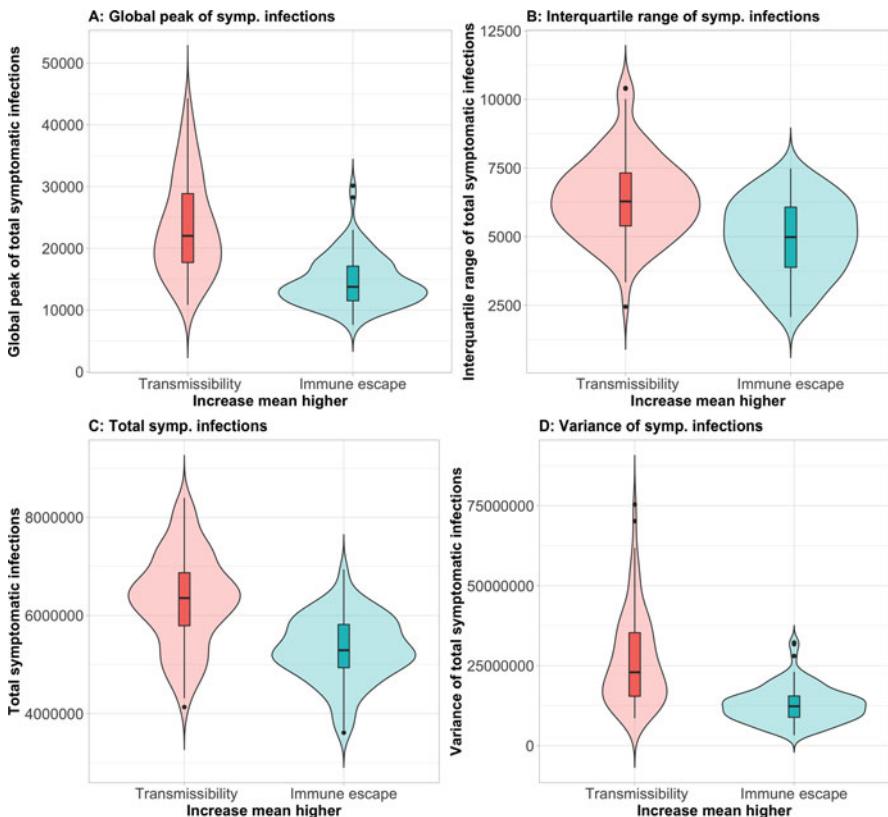


Fig. 11.6 Summary statistics for total symptomatic infections and peaks when the mean increase in transmission is greater than the mean increase in immune escape, and vice versa: (a) Global peak of symptomatic infections when the mean increase in transmission is greater than the mean of increase in immune escape capacity. (b) Interquartile range of symptomatic infections when the increase in transmission is higher than increase immune escape versus when the increase in immune escape is higher than the increase in transmission. (c) Total symptomatic infections for the two scenarios. (d) Variance of symptomatic infections under the scenarios for increase in transmission and immune escape, respectively. The variant arrival rate is fixed at 4 months. For higher increase in transmission we assume $\beta_{inc} = 35\%$, and $\epsilon_{inc} = 10\%$ for increase in immune escape. The values are switched when increase in immune escape is higher than increase in transmission. For consistency, the variances of the multivariate normal distributions are set at 0.0041 and 0.0021, respectively, and also switched accordingly when the means are switched. All other baseline parameters are kept as in Fig. 11.3

reduction in boosting will lead to higher peaks and higher total cases than continuing boosting. In most cases, a 90% reduction in the boosting rate leads to almost double the predicted symptomatic infections under the continuous boosting scenario (Fig. 11.7a–d). The outcomes are also more variable.

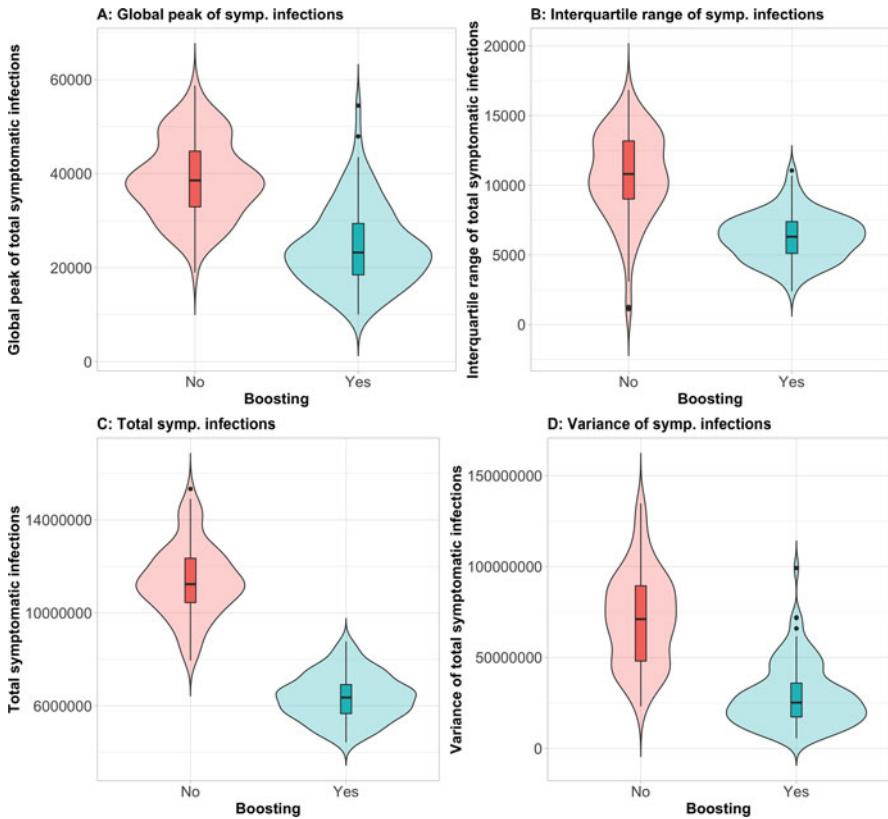


Fig. 11.7 Violin plots showing distribution of summary statistics for multiple runs of model simulation, comparing the impact, on symptomatic infections and peaks, of continuous boosting and discontinuing booting. **(a)** Global peaks of symptomatic infections for two scenarios—boosting and no boosting. **(b)** Two violin plots showing the interquartile range of total symptomatic infections over a period of 3 years, for boosting and no boosting scenarios. **(c)** Total symptomatic infections for the two scenarios **(d)**. The variance of total symptomatic infections for the two scenarios. Here $\beta_{\text{inc}} = 35\%$, $\epsilon_{\text{inc}} = 10\%$ for the red violins, and the means and variances are reversed for the blue violins. Baseline parameters are kept as in Fig. 11.3, and setting variants arrival rates to 4 months and boosting parameter b is assumed to be 0.0075 for boosting and 0.00075 for the low-boosting scenario

11.5 Discussion

Two of the key factors that determine whether a new variant will spread are the immune profile of the population and the growth advantage of the new variant over the currently circulating types. The former depends on the arrival time of the variant, as population immunity is dynamic [4] and changes depending on vaccination uptake and level of past infections, while the latter depends on the phenotype of the invading variant. Our hybrid model incorporates stochasticity in arrival times

and phenotypic traits, to gain insight on the long-term dynamics of COVID-19 in BC. We note that our long-term scenarios did not consider future implementation of strong physical distancing measures or other policy changes, or the development and deployment of new vaccines. Although some non-pharmaceutical measures such as mask mandates, mask encouragement, closures or guidelines could be implemented if cases were to rise rapidly, this seems unlikely at the time of writing, particularly as long as vaccine efficacy against hospitalization and severe disease remains high.

Our model assumes that there can be at most two concurrently circulating variants at any point in time. This is a simplifying assumption. In practice, many variants could be co-circulating. Our modelling approach implicitly assumes that selective sweeps limit diversity, which has frequently been the case in the first 2 years of the pandemic. Alternatively, in our model, each “strain” could be interpreted to describe a diverse collection of very similar co-circulating sub-strains. Furthermore, our assumption that the recovery rate is independent of the strain is also a limitation of the model, as evolution can impact the duration of infectiousness and the course of infection [16]. However, we do not expect these limitations to greatly impact the long-term dynamics of COVID-19.

Variants, whose increase in transmissibility is higher than the increase in immune evasion capacity, have more impact than the converse. This implies that when transmission is relatively low, even if the variant has higher escape capacity, the low transmission will limit the impact of the variant. Conversely, with a minimal increase in escape capacity in a variant with high transmission, the impact is multiplicative rather than additive. In a projection made for Omicron [44], the worst case scenario is observed when the variant is as transmissible as Delta and has higher immune escape capacity than Delta. Delta was already highly transmissible [11] and further increases in escape capacity can cause a substantial wave. We observe a similar trend in our model projection: when a variant combines high transmission and escape capacity, a high peaked wave is achieved. This suggests that major future waves are likely to have advantages in both transmissibility and immune escape capacity rather than showing improved competence in just one of the two traits.

Many jurisdictions have lifted most of their physical distancing measures [5] and now rely on booster doses to help control/avert resurgence of cases when new variants emerge. However, the uptake of booster doses may well drop over time. Our results show the negative consequences when the uptake of booster doses is reduced by 90% or suspended entirely. This suggests that continuous boosting may continue to be necessary until sterilizing vaccines with long-lasting immunity are available and widely used. Moreover, effective booster shots and vaccinations have an impact not only on the likelihood of infection in susceptible individuals upon contact but also on the infectiousness of vaccinated individuals [25]. If vaccines are effective in both regards, the likelihood of transmission within vaccinated populations is further reduced.

The necessity of boosting and of otherwise reducing infections depends on the consequences of infection. Here, we have not modelled evolution of the intrinsic severity of SARS-CoV-2 nor projected health-care burden. Since severe disease

occurs after the opportunity for transmission has largely passed, selection does not particularly favour increased or decreased severity directly [9, 31], and while Omicron is intrinsically milder than its predecessors, both alpha and delta are more severe than theirs [13]. Immunity against severe disease is longer-lasting than immunity against infection [24], and indeed protection against severe disease means that we can tolerate much higher levels of infection without exceeding health-care capacity than we could in the first part of the pandemic. However, infection can cause harm beyond acute disease (hospitalization, intensive care requirements) and death, and we are still discovering more of the consequences of even mild SARS-CoV-2 infection day by day [15, 20, 34, 37], and clinically mild infection may still result in lost work, increases in care hours and other impacts. Accordingly, there are likely to be benefits to reducing the burden of infection, and regular boosting is an available tool for this aim.

11.6 Conclusion

This modelling framework provides a tool for simulating the medium-term impact of possible future variants of SARS-CoV-2. We found that under the assumptions in our model, if new variants keep emerging at the current rate, stable endemic mode is not achieved within the next 3 years. We suggest that proactive measures should be deployed to reduce vaccine hesitancy and booster fatigue. The development and deployment of efficacious vaccines that confer broad, sterilizing immunity, or vaccines conferring longer-lasting immunity, are likely to be important tools in the longer term.

11.7 Supplementary Information

The model was used extensively to produce weekly COVID-19 case forecasts in Canada during the Omicron wave. The forecasts were done for the six largest provinces (with more than a million population): Alberta (AB), British Columbia (BC), Manitoba (MB), Ontario (ON), Quebec (QC) and Saskatchewan (SK), and we assessed the impact of relaxation of public health measures on numbers of “reportable” cases.

The provincial forecasts are further combined to make a Canada-wide forecast. The weekly forecasts were shared regularly with collaborators in the Public Health Agency of Canada (PHAC). We included samples of the forecasts as model calibration and to demonstrate that the model generates realistic case trajectories for multiple Canadian jurisdictions (Figs. 11.8, 11.9, and 11.10).

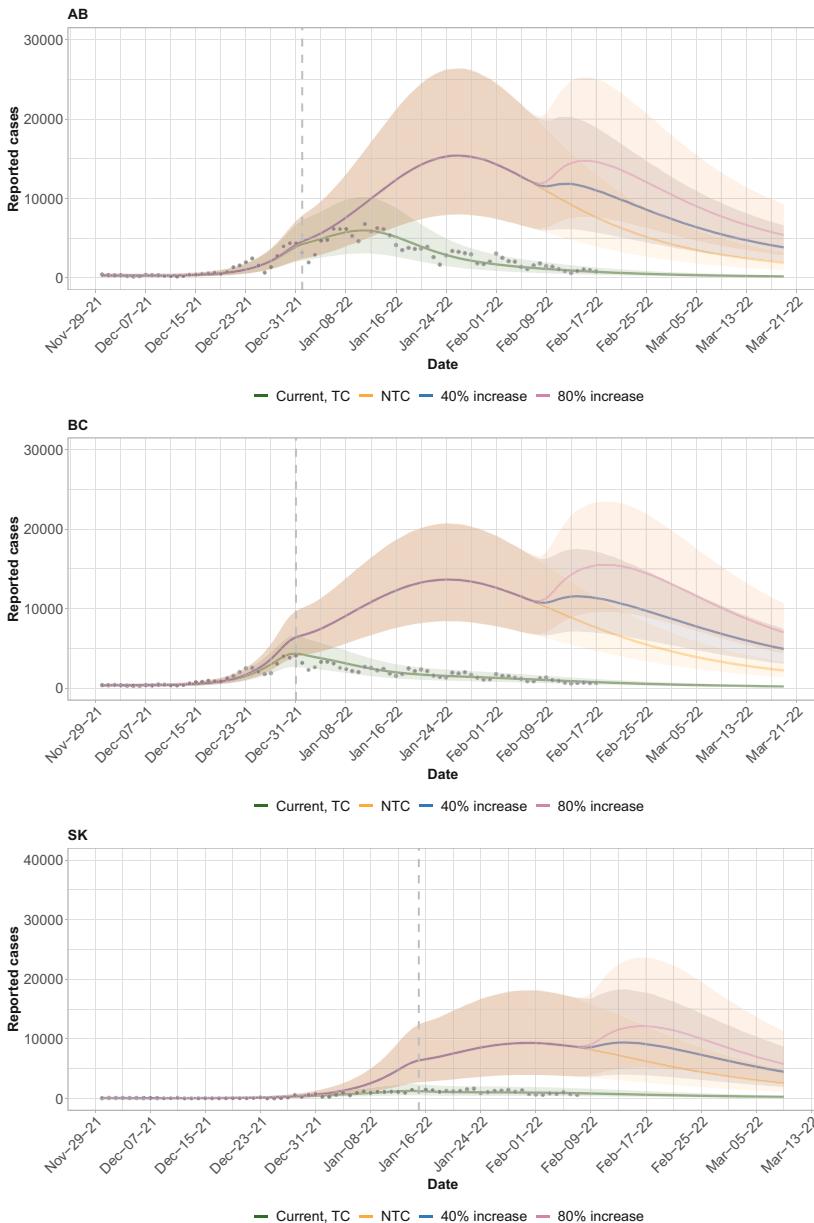


Fig. 11.8 Daily COVID-19 forecasts for Ontario, Quebec and Manitoba from February to March 2022. The “Current” (green line) projection shows model fit to reported cases factoring in under-reporting of cases due to testing constraints. The orange line (NTC) shows reportable cases assuming there are no testing constraints. Blue and purple lines show projections when public health measures are relaxed leading to 40% and 80% increase in transmission, respectively. The black dots show reported cases and the ribbons show 2.5% and 97.5% quantiles

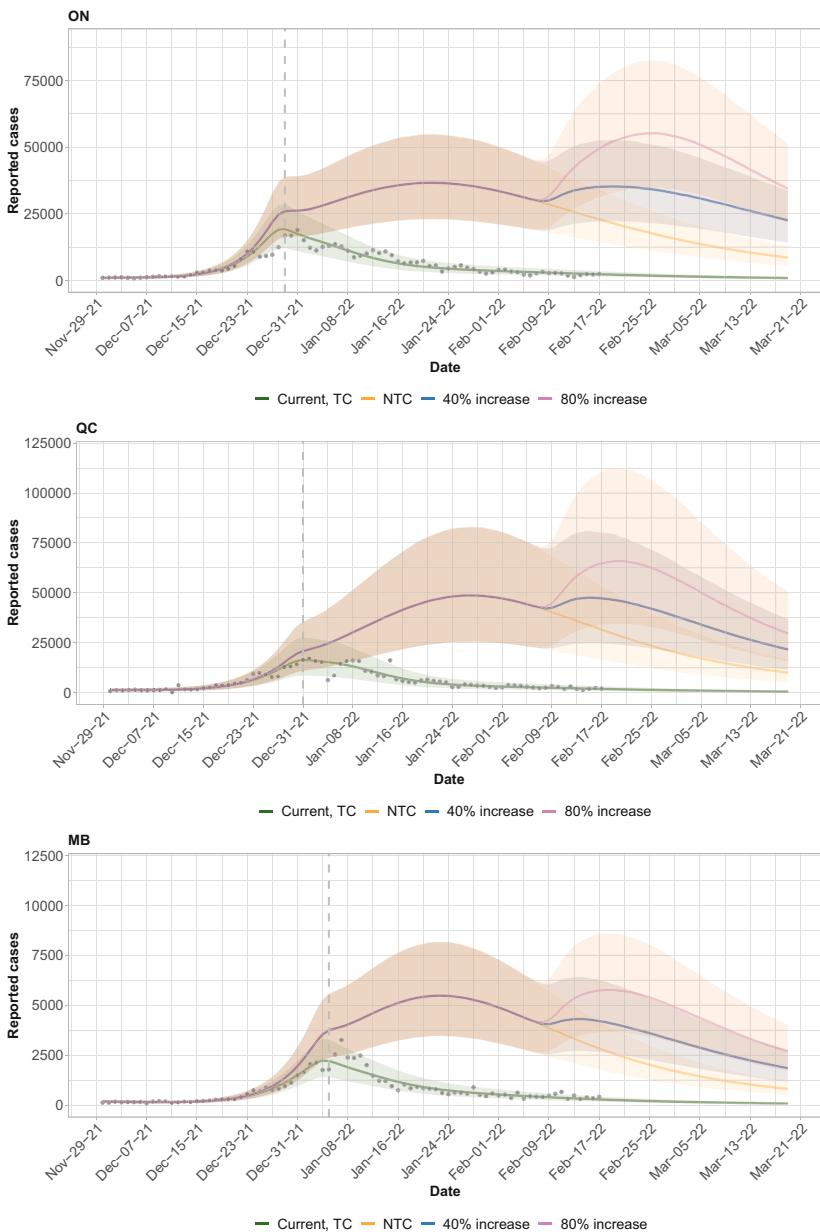


Fig. 11.9 Daily COVID-19 forecasts for Ontario, Quebec and Manitoba from February to March 2022. The “Current” (green line) projection shows model fit to reported incidence with under-reporting of cases due to limited testing capacity. The black dots show reported cases and the ribbons show 2.5% and 97.5% quantiles. The orange line (NTC) shows reportable cases assuming testing resources were not limited. Blue and purple lines show projections when public health measures are relaxed leading to 40% and 80% increase in transmission, respectively

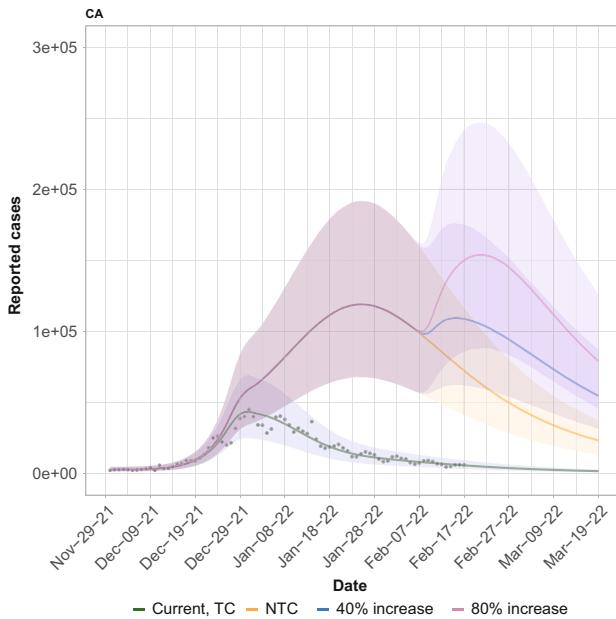


Fig. 11.10 Combined daily COVID-19 forecasts for Canada from February to March 2022. Similar to the provincial forecasts “Current” (green line) projection shows model fit to reported incidence with under-reporting of cases due to limited testing capacity. The grey dots show reported case numbers for the six provinces and the ribbons show 2.5% and 97.5% quantiles. The orange line (NTC) shows reportable cases when testing resources are unlimited. Blue and purple lines show projections when public health measures are relaxed leading to 40% and 80% increase in transmission, respectively

References

1. Are, E.B., Song, Y., Stockdale, J.E., Tupper, P., Colijn, C.: Covid-19 endgame: from pandemic to endemic? Vaccination, reopening and evolution in a well-vaccinated population (2021). medRxiv
2. Badr, H.S., Du, H., Marshall, M., Dong, E., Squire, M.M., Gardner, L.M.: Association between mobility patterns and covid-19 transmission in the USA: a mathematical modelling study. Lancet Infect. Dis. **20**(11), 1247–1254 (2020)
3. Baker, R.E., Park, S.W., Yang, W., Vecchi, G.A., Metcalf, C.J.E., Grenfell, B.T.: The impact of covid-19 nonpharmaceutical interventions on the future dynamics of endemic infections. Proc. Natl. Acad. Sci. **117**(48), 30547–30553 (2020)
4. Barker, P., Hartley, D., Beck, A.F., Oliver, G.H., Sampath, B., Roderick, T., Miff, S.: Rethinking herd immunity: managing the covid-19 pandemic in a dynamic biological and behavioral environment. NEJM Catal. Innovations Care Delivery **2**(5) (2021)
5. Bocean, C.G., Puiu, S., Vărzaru, A.A.: Paradigm shifting—the use of mobile communications at work and the subsequent effects. Electronics **10**(22), 2747 (2021)
6. British Columbia Center for Diseases Control (BCCDC) and Provincial Health Service Authority (PHSA) covid-19 dashboard (2022). <https://experience.arcgis.com/experience/a6f23959a8b14bfa989e3cda29297ded>

7. Cheung, C., Lyons, J., Madsen, B., Miller, S., Sheikh, S.: The bank of Canada covid-19 stringency index: measuring policy response across provinces. Tech. rep., Bank of Canada (2021)
8. Cole, S.R., Chu, H., Greenland, S.: Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *Am. J. Epidemiol.* **179**(2), 252–260 (2014)
9. Day, T., Gandon, S., Lion, S., Otto, S.P.: On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* **30**(15), R849–R857 (2020)
10. Dyson, L., Hill, E.M., Moore, S., Curran-Sebastian, J., Tildesley, M.J., Lythgoe, K.A., House, T., Pellis, L., Keeling, M.J.: Possible future waves of SARS-CoV-2 infection generated by variants of concern with a range of characteristics. *Nat. Commun.* **12**(1), 1–13 (2021)
11. Earnest, R., Uddin, R., Matluk, N., Renzette, N., Turbett, S.E., Siddle, K.J., Loreth, C., Adams, G., Tomkins-Tinch, C.H., Petrone, M.E., et al.: Comparative transmissibility of SARS-CoV-2 variants delta and alpha in new England, USA. *Cell Rep. Med.* **3**(4), 100583 (2022)
12. Endo, A., Abbott, S., Kucharski, A.J., Funk, S., et al.: Estimating the overdispersion in covid-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5** (2020)
13. Fisman, D.N., Tuite, A.R.: Evaluation of the relative virulence of novel SARS-CoV-2 variants: a retrospective cohort study in Ontario, Canada. *CMAJ* **193**(42), E1619–E1625 (2021)
14. Götz, T., Bock, W., Rockenfeller, R., Schäfer, M.: A two-strain SARS-CoV-2 model for Germany—evidence from a linearization (2021). arXiv preprint arXiv:2102.11333
15. Hastie, C.E., Lowe, D.J., McAuley, A., Winter, A.J., Mills, N.L., Black, C., Scott, J.T., O'Donnell, C.A., Blane, D.N., Browne, S., Ibbotson, T.R., Pell, J.P.: Outcomes among confirmed cases and a matched comparison group in the Long-COVID in Scotland study. *Nat. Commun.* **13**(1), 5663 (2022)
16. Hay, J.A., Kissler, S.M., Fauver, J.R., Mack, C., Tai, C.G., Samant, R.M., Connelly, S., Anderson, D.J., Khullar, G., MacKay, M., et al.: Viral dynamics and duration of PCR positivity of the SARS-CoV-2 omicron variant (2022). MedRxiv
17. He, D., Artzy-Randrup, Y., Musa, S.S., Gräf, T., Naveca, F., Stone, L.: The unexpected dynamics of covid-19 in Manaus, Brazil: was herd immunity achieved? MedRxiv (2021)
18. Huang, D., Tao, H., Wu, Q., Huang, S.Y., Xiao, Y.: Modeling of the long-term epidemic dynamics of covid-19 in the united states. *Int. J. Environ. Res. Public Health* **18**(14), 7594 (2021)
19. Kamiya, T., Alvarez-Iglesias, A., Ferguson, J., Murphy, S., Sofonea, M.T., Fitz-Simon, N.: Estimating time-dependent infectious contact: a multi-strain epidemiological model of SARS-CoV-2 on the island of Ireland (2022). medRxiv
20. Kompaniyets, L., Bull-Otterson, L., Boehmer, T.K., Baca, S., Alvarez, P., Hong, K., Hsu, J., Harris, A.M., Gundlapalli, A.V., Saydah, S.: Post-COVID-19 Symptoms and conditions among children and adolescents—United States, March 1, 2020–January 31, 2022. *MMWR Morb. Mortal. Wkly. Rep.* **71**(31), 993–999 (2022)
21. Kucharski, A.J., Russell, T.W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R.M., Sun, F., Jit, M., Munday, J.D., et al.: Early dynamics of transmission and control of covid-19: a mathematical modelling study. *Lancet Infect. Dis.* **20**(5), 553–558 (2020)
22. Ledford, H., et al.: How severe are omicron infections. *Nature* **600**(7890), 577–578 (2021)
23. Lima, C.M.A.d.O.: Information about the new coronavirus disease (covid-19) (2020)
24. Lin, D.Y., Gu, Y., Xu, Y., Wheeler, B., Young, H., Sunny, S.K., Moore, Z., Zeng, D.: Association of primary and booster vaccination and prior infection with SARS-CoV-2 infection and severe covid-19 outcomes. *JAMA* **328**(14), 1415–1426 (2022)
25. Lipsitch, M., Dean, N.E.: Understanding covid-19 vaccine efficacy. *Science* **370**(6518), 763–765 (2020)
26. Liu, J., Liu, Y., Xia, H., Zou, J., Weaver, S.C., Swanson, K.A., Cai, H., Cutler, M., Cooper, D., Muik, A., et al.: Bnt162b2-elicted neutralization of b. 1.617 and other SARS-CoV-2 variants. *Nature* **596**(7871), 273–275 (2021)
27. Lynge, F.P., Kirkeby, C.T., Denwood, M., Christiansen, L.E., Mølbak, K., Møller, C.H., Skov, R.L., Krause, T.G., Rasmussen, M., Sieber, R.N., et al.: Transmission of SARS-CoV-2 omicron variants ba. 1 and ba. 2: evidence from Danish households (2022). MedRxiv

28. Mahase, E.: Delta variant: What is happening with transmission, hospital admissions, and restrictions? (2021)
29. Mazzucco, S., Campostrini, S.: Life expectancy drop in 2020. estimates based on human mortality database. *PLoS One* **17**(1), e0262846 (2022)
30. Mohsin, M., Mahmud, S.: Omicron SARS-CoV-2 variant of concern: A review on its transmissibility, immune evasion, reinfection, and severity. *Medicine* **101**(19), e29165 (2022)
31. Otto, S.P., Day, T., Arino, J., Colijn, C., Dushoff, J., Li, M., Mechai, S., Van Domselaar, G., Wu, J., Earn, D.J.D., Ogden, N.H.: The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr. Biol.* (2021)
32. Park, K., Choi, S.J., Shin, E.C.: Omicron subvariants, including ba. 4 and ba. 5, substantially preserve t cell epitopes of ancestral SARS-CoV-2. *Immune Netw.* **31**(14), R918–R929 (2021)
33. Parker, M.R., Li, Y., Elliott, L.T., Ma, J., Cowen, L.L.: Under-reporting of covid-19 in the northern health authority region of British Columbia. *Can. J. Stat.* **49**(4), 1018–1038 (2021)
34. Phetsouphanh, C., Darley, D.R., Wilson, D.B., Howe, A., Munier, C.M.L., Patel, S.K., Juno, J.A., Burrell, L.M., Kent, S.J., Dore, G.J., Kelleher, A.D., Matthews, G.V.: Immunological dysfunction persists for 8 months following initial mild-to-moderate SARS-CoV-2 infection. *Nat. Immunol.* **23**(2), 210–216 (2022)
35. Port, J., Yinda, C.K., Avanzato, V., Schulz, J., Holbrook, M., van Doremalen, N., Shaia, C., Fischer, R., Munster, V.: Increased aerosol transmission for b. 1.1. 7 (alpha variant) over lineage a variant of SARS-CoV-2. Research Square (2021)
36. Pulliam, J.R., van Schalkwyk, C., Govender, N., von Gottberg, A., Cohen, C., Groome, M.J., Dushoff, J., Mlisana, K., Moultrie, H.: Increased risk of SARS-CoV-2 reinfection associated with emergence of omicron in South Africa. *Science* **376**(6593), eabn4947 (2022)
37. Raisi-Estabragh, Z., Cooper, J., Salih, A., Raman, B., Lee, A.M., Neubauer, S., Harvey, N.C., Petersen, S.E.: Cardiovascular disease and mortality sequelae of COVID-19 in the UK Biobank. *Heart* (2022)
38. Sigal, A., Milo, R., Jassat, W.: Estimating disease severity of omicron and delta SARS-CoV-2 infections. *Nat. Rev. Immunol.* **22**(5), 267–269 (2022)
39. Su, Z., Cheshmehzangi, A., McDonnell, D., da Veiga, C.P., Xiang, Y.T.: Mind the “vaccine fatigue”. *Front. Immunol.* **13**, 839433 (2022)
40. Sun, J., Chen, X., Zhang, Z., Lai, S., Zhao, B., Liu, H., Wang, S., Huan, W., Zhao, R., Ng, M.T.A., et al.: Forecasting the long-term trend of covid-19 epidemic using a dynamic model. *Sci. Rep.* **10**(1), 1–10 (2020)
41. Tan, W., Wong, L., Leo, Y., Toh, M.: Does incubation period of covid-19 vary with age? A study of epidemiologically linked cases in Singapore. *Epidemiol. Infect.* **148**, e197 (2020)
42. Tuite, A.R., Fisman, D.N., Greer, A.L.: Mathematical modelling of covid-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ* **192**(19), E497–E505 (2020)
43. Xia, S., Wang, L., Zhu, Y., Lu, L., Jiang, S.: Origin, virological features, immune evasion and intervention of SARS-CoV-2 omicron sublineages. *Signal Transduction Targeted Ther.* **7**(1), 1–7 (2022)
44. Zuo, C., Meng, Z., Zhu, F., Zheng, Y., Ling, Y.: Assessing vaccination prioritization strategies for covid-19 in South Africa based on age-specific compartment model. *Front. Public Health* **10**, 876551 (2022)