Journal Pre-proof

Big Data Technology in Infectious Diseases Modeling, Simulation and Prediction After the COVID-19 Outbreak: A Survey

Honghao Shi, Jingyuan Wang, Jiawei Cheng, Xiaopeng Qi,
Hanran Ji, Claudio J. Struchiner, Daniel A.M. Villela,
Eduard V. Karamov, Ali S. Turgiev

PII: S2667-1026(23)00003-7
DOI: https://doi.org/10.1016/j.imed.2023.01.002
Reference: IMED 67

To appear in: *Intelligent Medicine*

Received date: 5 August 2022
Revised date: 6 December 2022
Accepted date: 4 January 2023

Please cite this article as: Honghao Shi, Jingyuan Wang, Jiawei Cheng, Xiaopeng Qi, Hanran Ji, Claudio J. Struchiner, Daniel A.M. Villela, Eduard V. Karamov, Ali S. Turgiev, Big Data Technology in Infectious Diseases Modeling, Simulation and Prediction After the COVID-19 Outbreak: A Survey, *Intelligent Medicine* (2023), doi: https://doi.org/10.1016/j.imed.2023.01.002

Big Data Technology in Infectious Diseases Modeling, Simulation and Prediction After the COVID-19 Outbreak: A Survey

$HonghaoShi^{1}, JingyuanWang^{1,*}, JiaweiCheng^{1},$
$XiaopengQi^{2}, HanranJi^{2}, ClaudioJ.Struchiner^{3,4},$
$DanielA.M.Villela^{5}, EduardV.Karamov^{6,7}, AliS.Turgiev^{6,7}$

Abstract—After the outbreak of COVID-19, the interaction of infectious disease systems and social systems has challenged traditional infectious disease modeling methods. Starting from the research purpose and data, researchers improve the structure and data of the compartment model or use agents and AI-based models to solve epidemiological problems. In terms of modeling methods, the researchers use compartment subdivision, dynamic parameters, agent-based model methods, and AI-related methods. In terms of factors studied, the researchers studied 6 categories: human mobility, NPIs, ages, medical resources, human response, and vaccine. The researchers completed the study of factors through modeling methods, to quantitatively analyze the impact of social systems, and put forward their suggestions for the future transmission status of infectious diseases and prevention and control strategies. This review starts with a research structure of research purpose, factor, data, model, and conclusion, focusing on the post-COVID-19 infectious disease prediction simulation research, summarizes various improvement methods, and analyzes matching improvements for various specific research purposes.

Keywords—infectious disease model; data embedding; social system; dynamic; modeling the social systems; future preparedness

1 Introduction

Researchers use infectious disease models to study how infectious diseases spread, how fast they spread, and their spatial-temporal characteristics. The most commonly used infectious disease model is the population model represented by the compartment model, such as SIR[1], SEIR[2]. They divided the whole into several different groups according to different states, such as susceptible, infected, removed, etc., and used differential equations to define the mechanism of individual flow between groups.

From 1927, SEIR and other compartment models have been successfully applied to measles[3], SARS[4] and influenza A (H1N1)[5].

The three main elements of compartment models are Compartments, Transmissions(between compartments), and Parameters. After the outbreak of COVID-19, three elements of the compartment model faced limitations. These limitations stem from a common cause: the influence of social systems on infectious disease systems. Whether it's non-pharmacological interventions, vaccination strategies, the intensity of population activity, or the age distribution of those infected, these varied effects can each be attributed to a single factor in the social system. Researchers first determine the factor to be studied, then look for data that can characterize this factor, use modeling methods, and finally complete the research purpose through experiments.

In this review, Chapter 2 will introduce the basic concepts, application status, and limitations of infectious disease models. Chapter 3 categorizes the new modeling methods, and chapter4 categorizes the research factors, their corresponding data, and modeling methods. Chapter 5 would introduce the preparedness for future outbreaks.

- *. Correspondence:
  E-mail address: jywang@buaa.edu.cn;
  Full postal address: 37 Xueyuan Road, Haidian District, Beijing, P.R. China, 100191.
- 1. School of Computer Science and Engineering, Beihang University, Beijing, China
- 2. Center for Global Public Health, Chinese Center for Disease Control and Prevention, Beijing, China
- 3. Fundação Getúlio Vargas, Rio de Janeiro, Brazil
- 4. Instituto de Medicina Social Hesio Cordeiro, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil
- 5. Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil
- 6. Gamaleya National Research Center for Epidemiology and Microbiology of the Russian Ministry of Health
- 7. National Medical Research Center of Phthisiopulmonology and Infectious Diseases of the Russian Ministry of Health

Manuscript received July 29, 2022

2 Traditional Compartment Model

2.1 What is Compartment Model

A compartment model is a mathematical model that utilizes a set of compartments, parameters, and transformations to model the development of an infectious disease. The mathematical representation and practical significance of these three elements are shown in table 1:

Considering that all the compartments and parameters appear in the transformations in the form of a variable, a compartment model can be represented by a differential equation system.

Take the simplest SI model[6] as an example: SI contains two compartments: S(susceptible) and I(infectious), one pa-

TABLE 1
The mathematical representation and practical significance of compartments, parameters, and transformations

	mathematical representation	practical significance
Compartment	a state such as a susceptibility, infectious, or death	a function $F(t)$ relative to time, where t represents time, $F(t)$ represents the compartment's value at the corresponding time, the compartment's value represents the number of individuals in a corresponding state
Parameters	a numerical feature, such as infection rate, mortality rate, or time to onset	a figure
Transformation	the process of developing an epidemic, such as being infected and cured	a differential equation, where the left side is the differential of a certain compartment with respect to time t, and the right side is an expression consisting of compartment values, parameters, and constants

rameter: β for transmission rate, and two transformations. The differential equations of the SI model are:

$$\frac{dS}{dt} = -\frac{\beta \cdot S \cdot I}{N} \tag{2.1.1}$$

$$\frac{dI}{dt} = \frac{\beta \cdot S \cdot I}{N} \tag{2.1.2}$$

$$S + I \equiv N \tag{2.1.3}$$

2.2 SEIR Compartment Model

Before the outbreak of COVID-19 [2], the most widely used epidemic model was the SEIR compartment model. The differential equations of the SEIR model are:

$$\frac{dS}{dt} = -\frac{\beta \cdot S \cdot I}{N} \tag{2.2.1}$$

$$\frac{dE}{dt} = \frac{\beta \cdot S \cdot I}{N} - \alpha \cdot E \tag{2.2.2}$$

$$\frac{dI}{dt} = \alpha \cdot E - \gamma \cdot I \tag{2.2.3}$$

$$\frac{dR}{dt} = \gamma \cdot I \tag{2.2.4}$$

$$S + E + I + R \equiv N \tag{2.2.5}$$

In this system of equations, the meaning of each compartment, parameter, transformation, and constant N is shown in table 2 The SEIR model defines four states of the epidemic transmission process and the transformations between them. If it is possible to determine a set of initial compartments' values and parameters' values so that the compartments' values simulated by the model are consistent

TABLE 2
The meaning of SEIR compartment model

Type	mathematical representation	practical significance
Compartment	S	susceptible
	E	exposed
	I	infectious
	R	removed(death + recovered)
Parameter	β	transmission rate
	α	the reciprocal of latency
	γ	removal rate
Constant	N	population

with the real value time series, for example I(t) is consistent with the real number of existing patients, it means that the development process of the current infectious disease is consistent with the model. As a result, the model can be used for epidemic prediction.

2.3 How to use SEIR model to Predict Epidemic

2.3.1 The definition of Epidemic Prediction task

First, a definition should be made for the epidemic prediction task using the SEIR model:

- Through method f, **Parameters** are extracted from **Origin Data**.
- Through method h, **Initial Value** of compartments(as well as N) are determined from **Origin Data**.
- **Parameters** and **Initial Value** constitute **SEIR Input**.
- Through the principals of SEIR model, **SEIR Input** is calculated as **SEIR Output**.
- Through method g, **SEIR Output** forms the results of the task.

When researchers use the SEIR model for epidemic prediction, they first start from the original data, use the method f to extract the parameters, and use the method h to obtain the initial compartments' values and constant values. The initial compartments' values and constant values determine and only determine the input of the SEIR model. After the calculation of SEIR's differential equations, the researchers obtain the model output such as the time series of the number of infectious individuals and finally used another method g to extract the final result of the epidemic prediction task.

In practice, the original data should at least contain a time series of the number of infected and cured people and the population. In such a case, h and g are relatively simple:

- **Method h**: Taking the initial value of the number of infected people as the initial value of the I compartment, the sum of the initial value of cured and dead people as the initial value of the R compartment, and the population as N. Then estimating E compartment's initial value as $coe_E I$ times the initial value of I compartment [7] or the accumulated value of infected people in the next few days [8]. Finally, using $S + E + I + R \equiv N$ to calculate the initial value of the S compartment;

- **Method g**: Taking I compartment as the number of existing patients, and the sum of I and R compartment as the cumulative number of patients.

In fact, h and g vary substantially as researchers use external data to study new scenarios such as country-specific, population mobility, etc.[9], [10]. Therefore, the method f for extracting parameters would be introduced first.

2.3.2 How to extract Parameters

In different studies, the parameter extracting method f can be generally divided into three categories: apriori, operator calculation, and fitting. It is worth noting that different parameters in the same study may also be extracted in different ways. For example a study may use the apriori method to extract α, and the fitting method to extract β.

Apriori is to use information, facts, or data outside the extracting process of parameters, and directly extracting the value of the parameters. For example several studies use the reciprocal of the time of latency as the value of α[11], [12]. This time is derived from clinical data and has nothing to do with the SEIR model.

Operator calculation is to use of external data to calculate the value of parameters through several formulas or a simple mathematical model. For example, Kissler and Christine[13] used the strain-level incidence proxies and the generation interval distributions to estimate the daily effective reproduction number(R_u), then used R_u to determine parameters in SEIR model:

$$R_u = \sum_{t=u}^{u+i_{max}} \frac{b(t)g(t-u)}{\sum_{a=0}^{i_{max}} b(t-a)g(a)} \tag{2.3.1}$$

where b(t) is the strain-level incidence proxy on day t, g(a) is the value of the generation interval distribution at time a, and i_{max} is the maximum generation interval, set as the first day at which over 99of generation interval distribution had been captured[13].

Operator calculation is suitable for dealing with external factors that have little to do with the development scale and stage of infectious diseases, such as temperature, humidity, ultraviolet rays, and other climatic factors. Some studies from top journals have introduced the ERA5 dataset by invariably, and used the formula $\beta_t = exp\{a \cdot d_t + log(d_{max} - d_{min})\} + d_{min}$ to calculate β under the scenario of taking into account climate factors[14], [15], [16], where $\{d_t\}$ refers to temperature or UV rays from ERA5 dataset.

Fitting refers to calculating a set of optimal or relatively optimal parameters through an optimization method so that the simulated compartment values of this set of parameters are as close as possible to the real situation. The fitting method is the major method for extracting parameters, and it is also the only method whose results are strongly correlated with the actual development of the epidemic. For example, L Xue and others[17] used the MCMC method to extract the parameters in epidemic models, which modeled COVID-19 in Wuhan, Toronto, and Italy. Because the fitting method is based on real data, it is also called a data-driven method.

2.3.3 The weakness of Traditional Compartment Model

After the outbreak of COVID-19, the SEIR model has shown two limitations as an infectious disease model: it cannot modeling the real social systems and dynamics.

Although the SEIR model models the whole process of contact-exposure-onset-removal of the development of infectious diseases, it is too ideal for the assumptions of compartments and individuals. This is reflected in:

- Individuals in the same compartment are identical. For example, infected individuals transmit the disease to susceptible individuals at an average rate, and each individual has the same importance in the transmission chain.
- Each individual is indifferent without subjective initiative. Individuals will not change their action strategies or formulate non-pharmacological interventions(NPIs, similarly hereinafter) according to the development of the epidemic.
- The compartment is set according to the principle of the epidemic, not the actual observation data. For example, the infected person's compartment is set, but only the confirmed data can be obtained in reality, and the error of approximate substitution is unignorable.

With increased human mobility and the introduction of NPIs, the complex, dynamic spread of COVID-19 has diverged significantly from SEIR's single, static assumption. At the same time, the ability to obtain front-line data also limits the modeling capabilities of SEIR: Henrik Salje and others[18] modeled the different ages of individuals in the compartment and took into account the special propagation patterns of this scene in France, Vadim A. Karatayeva and others[19] modeled the dynamics of the transmission rate caused by dynamics in population mobility, and conducted a simulation experiment on the effect of NPIs based on this model.

3 Methods to Improve Traditional Compartment Model in COVID-19 era

The improvement about "Modeling the Dynamics" were well-known before COVID-19 outbreak, but using these improvement with big data to solve infectious disease modeling problem proliferate after COVID-19 outbreak. We review these methods here, and review specific research purpose with specific data in chapter 4.

3.1 Modeling the Dynamics

Dynamics means that the epidemic model varies with external factors such as spatial factors, temporal factors, and characteristics. According to the scale of variation, the methods of modeling dynamics can be divided into two categories: multi-stage models and parameter dynamics. Multi-stage models vary more but require more complete data and logic to support. Dynamic parameters only change the parameters of the model, which is more feasible under the premise of reasonable model design. In practical research, multi-stage models are often used to review specific epidemics, while parameter dynamics are suitable for extensive

research such as data analysis, simulation, prediction, and regression. The difference between the multi-stage model and dynamic parameters could be seen in table 3.

TABLE 3
The difference between multi-stage model and dynamic parameters

Aspect	multi-stage model	dynamic parameters
scale of variation	whole model	only parameters
required data	lots and complex	simple arrays usually, sometimes geospatial or graph data
research focus	when to divide stage, how to model each stage	how to use array data to make parameters dynamic
research interest	review specific epidemics	scientific purpose
popularity	low, only a bit	high, most of the research about modeling dynamics

3.1.1 Multi-stage Models

The multi-stage model refers to the use of different models to model the epidemic according to certain rules. The dynamic nature of the multi-stage model is usually reflected in the time dimension, which is derived from the NPIs, an external factor with great influence.

Xingjie Hao and Chaolong Wang[20] published their academic research on the review of COVID-19 in Wuhan in Nature. Using the SAPHIRE framework, they differentiated between symptomatic and asymptomatic infections and added presymptomatic compartments between exposed and infected compartments. Using this framework, they introduced the release timing of NPIs such as Wuhan's city closure and established a five-stage SEPIR-class epidemic model. The transformations of the model for each stage are the same, but the compartments' values are reset according to the end of the previous stage and the real data. The parameters are also completely re-fitted by the Markov chain Monte Carlo method. The cut-off points of the five stages are determined a priori by NPIs, such as 2020.1.22 (Wuhan's city closure) or 2022.2.2 (with the addition of clinical diagnosis criteria), and have nothing to do with confirmed case data or the SEPIR model itself.

A study published in PNAS by Daniel Duque and others[21] on social distancing and COVID-19 hospital surges also used a multi-stage model. Unlike Chaolong Wang, they presented a strategy for triggering short-term shelter-in-place orders when hospital admissions surpass a threshold. In other words, they use a certain data indicator to automatically obtain cut-off points, rather than manually specifying them based on external factors.

In the study of modeling dynamics using multi-stage models, manual formulation and automatic acquisition according to a certain index are two types of methods to obtain cut-off points. The former, such as the review of an epidemic in Wuhan by Wang[20] and the research on how Chinese NPIs control COVID-19 by maier[22], deal with irregular, sudden external influences like NPIs. The latter, such as the research on hospital surges by Duque[21], deal with external factors with certain regularity. It's usually a threshold like the hospital admissions threshold or the government response index threshold[23]. The difference between them is summarized in table 4

TABLE 4
The summary of multi-stage model

Aspect	Manuel	Automatic
basis	external human intervention	external indicator with a certain threshold
scenarios	NPIs	extensive external factors

At last, models in different stages are re-fitted as least, which is a variation on a larger scale compared to dynamic parameters.

3.1.2 Dynamic Parameters

Dynamic Parameters refer to the use of external data to change the parameters of the infectious disease model from a figure to an array or even a matrix by means of apriori, operators, or complex sub-models. Dynamic parameters are more of a modeling basis than an improvement. Dynamic parameters are to the COVID-19 model what the E compartment is to the SEIR model, a common feature of models. The ultimate purpose of dynamic parameters is very broad, but the direct reason is that the parameters of the model are affected by external factors, so as to study the relationship between those external factors and the epidemic. Compared with multi-stage models, research with dynamic parameters as the main method not only focuses on the review and analysis of past epidemics, but also on simulation experiments using models, and scientific conclusions are drawn from them. For example, Serina Chang and others[24] divided the places where people contact each other and cause infection into CBG (census block group) and POI (points of interest), where CBG is a geographic unit with a population of 600-3000, and POIs are frequented by people Places such as restaurants, grocery stores, and places of worship. The transmission rate β_{base} of all CBG home infections is the same value, while those of POIs are:

$$\beta_{p_j}^{(t)} = \varphi d_{p_j}^2 \frac{V_{p_j}^{(t)}}{a_{p_j}} \tag{3.1.1}$$

where $\varphi$is a same propagation constant for all POIs, a_{p_j} is the actual area of p_j, $V_{p_j}^{(t)}$is the volume of visitors at time t and $d_{p_j}$is the average hourly percentage of visitors visiting this POIs at any time. In this case, the transmission rate β is dynamic in the spatial dimension.

As defined in the extraction parameter section, researchers can also obtain dynamic parameters through apriori and operator calculation methods. In Chang's research[24], they use multiplication and division operators to introduce the relevant data of visitors to complete the parameter dynamic process. In addition, researchers will also use models in the fields of mathematics and computer science, such as Bayesian regression models, RNN-

class models, and GNN-class models, to extract dynamic parameters, that is, complex sub-model methods.

Compared with original infectious disease modeling, researchers introducing complex sub-models are more inclined to explore how to design sub-models to allow SEIR-like epidemic models, or more broadly, dynamic models(refers to models who model a dynamic system like an epidemic, "dynamic" here has a different meaning with those in "dynamic parameters"), to model the link between social systems and infectious disease systems. For example, Chang and others[25] discussed in detail the impact of NI(no interventions), CI(case isolation), HQ(home quarantine), SC(school closures) policies and their combinations on SD(social distance), and the spread of the epidemic. Among them, the results of SD are used to extract dynamic parameters in the SEIR model.

In the field of computer science, machine learning models are particularly suitable for the task of processing given data and discovering valuable information or conclusions. In research using complex sub-models to obtain dynamic parameters, most researchers use machine learning models instead of complex but traditional mathematical models. In fact, with the introduction of machine learning methods, the original SEIR's compartment settings and transformations have also undergone different changes.

For instance, Amray Schwabe and others[26] presented a paper at the KDD conference, using machine learning models to process massive mobile data, and use it to obtain dynamic parameters for infectious disease models. They designed the M2H(mobility to Hawkes process) model: used external data such as case data and mobile data to complete the fitting and application of the Hawkes process, and verified that its recurrence and epidemic prediction results were better than the SEIR model.

3.2 Modeling the Real Social Systems

Compared with modeling dynamics, modeling the real social systems covers a wider range of studies and can use more methods. If modeling dynamics mainly changes the parameters in the three elements of the compartment model, then modeling the real scene requires designing the settings and transformations between the compartments. From the perspective of modeling methods, the means of modeling real scenes can be divided into 2 categories: compartment subdivision and meta-population models.

3.2.1 Subdivided Models

Subdivided Models refer to further dividing one or more compartments in SEIR according to certain rules. From the perspective of mathematical modeling, compartment subdivision is to change a compartment C to C_1, C_2,..., C_n, and supplement the matching transformations. From the perspective of infectious disease modeling, compartment subdivision can be divided into two categories: horizontal and vertical. The former solves the difference between different individuals in the compartment, and the latter solves the problem that SEIR's compartment does not match the real data. "Horizontal" and "Vertical" describe their subdivision directions in the schematic.

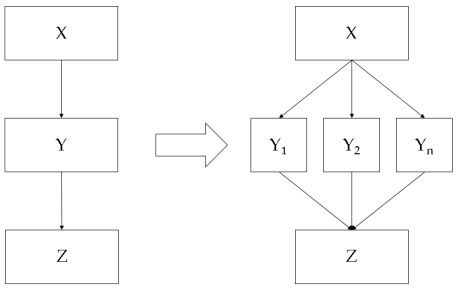

Fig. 1. schematic of horizontal subdivision

The schematic of the horizontal subdivision is shown in figure 1.

In the system of differential equations of the compartment model, the horizontal subdivision has the form of formula (3.2.1)-(3.2.11):

$$before: \tag{3.2.1}$$

$$\frac{dX}{dt} = -v_{xy} \tag{3.2.2}$$

$$\frac{dY}{dt} = v_{xy} - v_{yz} \tag{3.2.3}$$

$$\frac{dZ}{dt} = v_{yz} \tag{3.2.4}$$

$$X + Y + Z \equiv N \tag{3.2.5}$$

$$after: \tag{3.2.6}$$

$$\frac{dX}{dt} = -\sum_{i=1}^{n} v_{xy_i} \tag{3.2.7}$$

$$\frac{dY_i}{dt} = v_{xy_i} - v_{y_i z} \tag{3.2.8}$$

$$\frac{dZ}{dt} = \sum_{i=1}^{n} v_{y_i z} \tag{3.2.9}$$

$$X + \sum_{i=1}^{n} Y_i + Z \equiv N \tag{3.2.10}$$

(3.2.11)

In the horizontal subdivision, the subdivided compartments are "equal in status", and there is no transmission between these compartments. Therefore, the horizontal subdivision method is suitable for dealing with external factors, which are not related to epidemiological status, but can affect the development of infectious diseases, such as age[27], [28], [29], occupation[30] and work intensity[31]. For example Alessio Andronico and others[27] emphasized the impact of different age structures on hospitalizations by looking at histograms comparing the age distribution of hospitalized cases in French metropolis and French Guiana. And through the differences in the proportion of hospitalizations and deaths between the two, they further pointed out that considering the age structure of the population necessary, these differences were successfully predicted by the model used. Andronico subdivides the compartments horizontally into 8 age groups according to age, each coverint 10 years.

The schematic of the vertical subdivision is shown in figure 2.

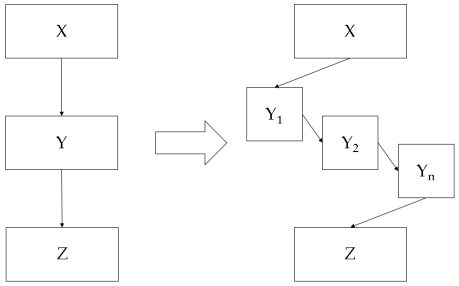

Fig. 2. schematic of vertical subdivision

In the system of differential equations of the compartment model, the vertical subdivision has the form of formula (3.2.12)-(3.2.24):

$$before: \tag{3.2.12}$$
$$\frac{dX}{dt} = -v_{xy} \tag{3.2.13}$$
$$\frac{dY}{dt} = v_{xy} - v_{yz} \tag{3.2.14}$$
$$\frac{dZ}{dt} = v_{yz} \tag{3.2.15}$$
$$X + Y + Z \equiv N \tag{3.2.16}$$
$$after: \tag{3.2.17}$$
$$\frac{dX}{dt} = -v_{xy_1} \tag{3.2.18}$$
$$\frac{dY_1}{dt} = v_{xy_1} - v_{y_1y_2} \tag{3.2.19}$$
$$\frac{dY_i}{dt} = v_{y_{i-1}y_i} - v_{y_iy_{i+1}}, where\ 1 < i < n \tag{3.2.20}$$
$$\frac{dY_n}{dt} = v_{y_{n-1}y_n} - v_{y_nz} \tag{3.2.21}$$
$$\frac{dZ}{dt} = v_{y_nz} \tag{3.2.22}$$
$$X + \sum_{i=1}^{n} Y_i + Z \equiv N \tag{3.2.23}$$
$$\tag{3.2.24}$$

The compartments after vertical subdivision can be transformed in sequence. Therefore, vertical subdivision methods are suitable for modeling more detailed epidemiological states such as presymptomatic, confirmed, isolated, and hospitalization [21]. In fact, the process of extending the SI model to the SEIR model can also be regarded as a vertical subdivision, that is, the I compartment is subdivided into E-I-R sub-compartments. For example, when discussing the consequences of the relaxation of school control measures in France, Laura Di Domenico and others [32] proposed an SEIR-based compartment model, which subdivided the follow-up status of severely infected patients vertically according to treatment methods and medical conditions, including hospitalization and ICU treat. Next, they can use data from hospital admissions and ICU admissions to refine the model, rather than uniformly treating them as infected people and placing them in the I compartment.

The difference between the multi-stage model and dynamic parameters could be seen in table 5.

TABLE 5
The summary of horizontal and vertical subdivision

Aspect	horizontal subdivision	vertical subdivision
purpose	solves the difference between individuals	matches the real data
relationship between sub-compartments	None	can be transformed in sequence
subdividing standard	not related to epidemiological status	related to epidemiological status

It is worth noting that horizontal and vertical segmentation methods are not contradictory. For research purposes, researchers usually use a combination of the two. For example Osmar Pinto Neto and others [7] proposed a SUEIHCDR model that utilizes sophisticated compartment subdivision methods. They add 4 compartments on the basis of SEIR: U(unsusceptible), H(hospitalized), C(critical), and D(death). Starting from the SI model, they did the following:

- Subdividing I into U and I horizontally, and the less(or even un-) susceptible people are modeled.
- Subdividing I into E-I-R vertically, as the SEIR model does. The R here refers specifically to recovered rather than removed.
- Subdividing R into R and H horizontally, modeling the population requiring hospitalization. The meaning of the transformation from I to R has also changed from a broad cure to a specific natural cure.
- Subdividing H into H, C, and D vertically, modeling the progression of hospitalized patients with progressive deterioration and eventual death.
- Complement the H-R, C-H transformation to model the recovery process of hospitalized or critically patients.

3.2.2 *Meta-population Models*

After the limitations of individual differences and data mismatches were addressed by subdivided models, the researchers go on to employ meta-population models to more accurately model the process of "transmission" [33], [34]. In the SEIR model, "transmission" occurs uniformly between susceptible and infected compartments. But actually, in human society, this process goes along social networks. The meta-population model regards each node on the social network as a population, and the transmission of infectious diseases within the population is uniform, which is in line with the SEIR model. Between populations, along every edge of a social network, there is a flow of individuals. After obtaining data on flows and policies (eg: not allowing the movement of infected patients), a meta-population model is built to model infectious diseases on this social network.

A general meta-population model differential equation system is as follows:

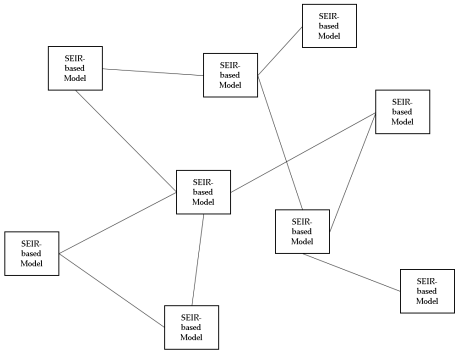

Fig. 3. schematic of meta-population model

$$\frac{d\boldsymbol{S}}{dt} = -\frac{multiply(\boldsymbol{S}, \boldsymbol{\beta} \cdot \boldsymbol{I})}{N} \tag{3.2.25}$$

$$\boldsymbol{S} = (S_1, S_2, ..., S_n)^T \tag{3.2.26}$$

$$\boldsymbol{\beta} = (\beta_{ij})_{n*n} \tag{3.2.27}$$

$$\boldsymbol{I} = (I_1, I_2, ..., I_n)^T \tag{3.2.28}$$

$$where \tag{3.2.29}$$

$$\boldsymbol{x} = (x_1, x_2, ..., x_n)^T \tag{3.2.30}$$

$$\boldsymbol{y} = (y_1, y_2, ..., y_n)^T \tag{3.2.31}$$

$$multipy(\boldsymbol{x}, \boldsymbol{y}) = (x_1 \cdot y_1, x_2 \cdot y_2, ..., x_n \cdot y_n)^T \tag{3.2.32}$$

Although $\boldsymbol{\beta}$ is used as the propagation matrix, the meta-population model only focuses on some specific elements or vectors, not a whole(and sometimes random) matrix. Write β_{ii} as the propagation coefficient within the i_{th} population, then $\boldsymbol{\beta_i} = (\beta_{i1}, \beta_{i2}, ..., \beta_{in})$ can be adapted into $\boldsymbol{\beta_i} = \beta_{ii} \cdot (c_{i1}, c_{i2}, ..., 1.0, ..., c_{in})$, and:

$$c_{ij} = 1.0, where\ i = j \tag{3.2.33}$$

$$c_{ij} = (mobility(i, j), where\ i \neq j \tag{3.2.34}$$

The schematic of the meta-population model is shown in figure 3.

The meta-population model provides an entrance for researchers to introduce Spatio-temporal big data, and epidemic simulation based on the propagation matrix and propagation map can achieve more refined quantitative results than the traditional compartment model. For example, the paper published on PNAS by Ruiyun Li and other researchers[14] simulated the development of the epidemic after closing several high-traffic populations(social network nodes), indicating that the high-traffic populations(nodes) were first subjected NPIs for 8 weeks, and then intervened others for 8 weeks can reduce the epidemic scale by 88%. The degree of effectiveness exceeds that of 12 weeks of NPIs on all populations, and the cost is less than the latter.

The use of meta-population models must be accompanied by dynamic parameters, and usually accompanied by compartment subdivision, that is, each population is a subdivided model instead of an SEIR model. In fact, in research on COVID-19, the model is usually a meta-population model where each population is a subdivided model, and the model has dynamic parameters[29], [35], [36], [37]. How multiple populations can be designed, compartments subdivided, and dynamic parameters extracted depends primarily on the purpose of the study and the data the researcher has. A good model can make the best use of data, obtain as much knowledge as possible, then help researchers model abstract influencing factors, and finally complete the research purpose.

3.3 Beyond Compartment Models - Agent-based Models

With the development of computer simulation technology and artificial intelligence technology, researchers have been able to introduce the agent-based model into the field of infectious disease modeling. The modeling object of the traditional model is the compartment, that is, the group. In order to reflect the differences within the group, the researchers subdivided the group. In this way, it is possible to study the factors that lead to such internal differences and their impact on the spread and development of the epidemic. But in subdivided models, even in a meta-population model, the modeling object is still the population, the group. The individuals inside each compartment are exactly the same, and only reflect the characteristics of the group.

In fact, when the amount of data used for model fitting remains unchanged, increasing the number of compartments without limitation will only reduce the prediction and simulation performance of the model. The agent-based model takes the individual as the modeling object and completes the modeling of the spread of infectious diseases, various medical interventions or NPIs, and individual subjective behavior by defining the state and behavior of the individual. The agent-based model usually simulates dozens or even tens of thousands of individuals at the same time. Through the statistics of individual states and behaviors, the research purpose can be accomplished[38], [39], [40].

In the application scenario of the agent-based model, the state categories of individuals are usually many and fine, and it is difficult to model by conventional subdivided models. In the infectious disease model, the behaviors of individuals usually include intrinsic behaviors (morbidity, death, cure, etc.) and extrinsic behaviors (still, moving, migrating, etc.), and the spread of infectious diseases is carried out, checked, and determined for all individuals in each time unit, and executed with a certain probability to individuals who meet the conditions. For example, Kim Sneppen and others published a paper on PNAS[41], using the agent model to complete the modeling of super communicators, and deploying super communicator agents in specific communication scenarios such as schools and workplaces to explore how to formulate epidemic prevention policies with the existence of super communicators. The research work published by Jesús A. Moreno López1 and others in Science[42] introduced electronic device tracking data to explore the intervention process of the COVID-19 epidemic, especially how age and heterogeneity in modeling parameter settings interact.

An agent-based model of the spread of COVID-19 has been developed by a consortium of Russian research centers. Since the development of an epidemic is a kind of a chain reaction, the authors viewed the simulation process as an analog of the method used to solve the neutron transport

equation for a heterogeneous medium in a multigroup approximation[43]. This model, demonstrating a good predictive power for metropolitan cities (Wuhan, New York, and Moscow), is currently being adapted to simulation of other viral respiratory infections and country-wide use.

On the basis of the above research, the researchers pushed the agent-based model from theory to application, and formed a series of agent-based model open source simulation platforms represented by covasim[44] and openABM-covid19[45]. These open-source platforms focus on multiple scenarios. They not only complete model design and simulation operation, but also parameterize engineering factors such as scenario characteristics, data analysis methods, and data visualization methods, and form a complete open-source code, homepage, published papers and user manual documentation website.

3.4 AI techniques in Improving COVID-19 Modeling

Researchers also use AI techniques to solve the problem of static and too ideal scenes of traditional models. This combination of compartment model and AI is a hybrid Physics-ML model[46]:

- Residual modeling,
- Output of physical model as input to ML model,
- Replacing part of a physical model with ML,
- Combining predictions from both physical model and ML model,
- ML informing or augmenting physics-model for inverse modeling.

In the scenario where AI technology and compartment model are combined to deal with infectious diseases, by examining how models are output and the application of the laws of physics, the first three are AI in the auxiliary position, and the input and output of the model are still completed by the compartment model. The latter two are AI in the dominant position, and the input and output are completed under the guidance of physical rules (the principle of the compartment model). In addition, there is a model that is made purely of AI methods and has nothing to do with the compartment model. We call these three types, in turn, Assisted, Coexistence, and Pure AI

- **Assisted**: The compartment model completes all the processes such as modeling, parameter calculation, simulation experiments, etc. AI methods are only used to process part of the external data to achieve the goal of dynamically changing compartment values or parameters.
- **Coexistence**: The designing, modeling, and simulation experiments are completed by the AI model. However, the principles of the compartment model are used, such as how the population in the cabin is divided, or the physical laws described by the dynamic equations.
- **Pure AI**: Those studies that have nothing to do with compartment models, but belong to the field of infectious disease modeling

In this review, we briefly discuss and analyze the first two categories.

3.4.1 AI Assist Compartment Model

Dynamic Parameter is the primary method of AI assisting compartment model. Compared with the numerical values or operators used in other dynamic parameter models, the method here uses a complete, complex, and independent AI model to obtain dynamic parameters and use them in the compartment model. Salah Ghamizi and others[47] published a paper in KDD'20, studying DN-SEIR, a data-driven approach to evaluating the effective reproduction number of the COVID-19 epidemic. They build an AI model(DNN) to predict the reproduction rate(Rt), then use Rt to activate SEIR model.

3.4.2 Coexistence of AI and Compartment Model

The knowledge of introducing kinetic equations into AI models falls under a larger scope of research: AI research with physical knowledge. In the field of infectious disease modeling, such models incorporate SEIR, or other variants of differential equations, into an AI model in the form of the loss function. Lijing Wang and others[48] used a Causal-based Graph Neural Network(CausalGNN) that learns spatiotemporal embedding in a latent space where graph input features and epidemiological context are combined via a mutually learning mechanism. In their model, the graph of disease dynamics is encoded through the time dimension. The encoding layers include feature encoding, spatiotemporal encoding, and finally causal encoding. At the top layer, encoding results are calculated in the SIRD compartment model, then SIRD's results are used as the input of the next time unit.

4 Data, Factors, and Modeling Methods of Research Interests

4.1 Data, Factors, Modeling Methods and Research Interests

At the end of the section "Meta-population Model", it is mentioned that *A good model can make the best use of data, obtain as much knowledge as possible, then help researchers model abstract influencing factors, and finally complete the research purpose*. In fact, after the outbreak of COVID-19, the research on epidemic modeling has been inseparable from data, factors, and models. All research today is not limited to infectious disease itself but uses external data and models to model certain or certain types of influencing factors, so as to carry out simulation experiments and draw conclusions. Data, factorsa and models will be referred to as DFM hereinafter. Besides data, factors, and models, models are the subject of exploiting data, at the core of the modeling factor, as a tool for research purposes, and the most complex part. Therefore, the classification, principles, and application cases of models have been introduced in detail in chapter 2. Here will be a brief introduction to data and factors.

4.1.1 Data

The data introduced in the infectious disease model can be divided into case data and non-case data according to content and source, and can be divided into point feature data, array data, and geospatial data according to the data format.

Case data mathematically includes exposure, infection, and removal data associated with the compartment, and in practice includes three categories: confirmed, cured, and dead. JHU(John Hopkins University)'s CSSE(Center for Systems Science and Engineering) is dedicated to building GIS(Geographic Information System) and collecting data, and after the outbreak of COVID-19, a database of cases from countries around the world and states in the United States has been established[49]. This database is accurate for every country, updated every day, and is extremely widely used. Related papers have been cited more than 6,900 times.

Besides, in order to complete the vertical subdivided model and dynamic parameters, researchers introduced a type of "detailed case data". In addition to the basic confirmed, cured, and death, it also includes critical illness, ICU, hospitalization, asymptomatic infection, close contact, nucleic acid testing, vaccination, etc. The case data is directly related to the compartment itself in content, mainly from the reports and epidemiological investigations of front-line staff. This type of data is mainly used as the true value to participate in the fitting process and is used as an index to compare with some results during the simulation experiments.

For example, Joshua S. Weitz[50] introduced a detailed US case dataset to complete his social system self-feedback model. That dataset includes 10 types of data: Positive, Negative, Hospitalization, ICU, Ventilator, Cure, Data Quality Rating (Confidence Level), Death, Data Recording Time, Floating Space, and each type has 3 statistical dimensions: Daily new, Existing, Accumulated.

By definition, non-case data includes any data that is not case data but is used by researchers to model infectious diseases. Therefore, the discussion of non-case data focuses on its data format. Each format represents a class of information, has similar processing operators, and is applied in the same way in the model.

Point feature data is a type of feature data with a key-value format. For each object, such as country, country, POI or even climate zone, has an ID and several dimension attributes. Such data are processed into feature vectors by traditional statistical methods and data mining methods, which are then used to obtain dynamic parameters based on characteristics or to build meta-population models. For example the mobility data between each CBGs and POIs can be described as a point feature dataset, where the objects are CBGs or POIs, and attributes are traffic flows, population, etc.[24]. Array data(or sequence data) is a type of data expanded in the time dimension. Each moment contains several attributes, which can be embedded into a feature vector. Sequence data is very common in case data, for example, daily confirmed data is sequence data. In non-case data, sequence data is mainly used to extract dynamic parameters. Methods for processing sequence data include regression, fitting, and RNN-class deep learning models for extracting sequence information.

Geospatial data is based on GIS and contains descriptions of geographic information. In addition to geographic information, each object also includes several attributes, which can also be embedded into a feature vector. For instance, Carleton and others[15] use geospatial data(ERA-5 dataset) about UV rays to complete their research about the influence of UV rays on the epidemic.

The summary of case and non-case data could be seen in table 6 and table 7.

TABLE 6
The difference between case and non-case data

Aspect	Case Data	Non-case Data
source	medical and epidemiological staff	everywhere
content	the number of individuals corresponding with compartment	everything
usage	as the true value	extracting parameters and building meta-population models

TABLE 7
The difference between 3 types of non-case data

Aspect	Point Feature Data	Array data	Geospatial data
format	key-value	array(sequence)	mostly .shapefile or .stata data
how to use	embedding and bringing into formula	regression, fitting, and RNN-class deep learning models	pre-processing and bring into formula

4.1.2 Factors

A factor is an abstract object that can be described in natural language and is a bridge between research interests and data. In order to achieve the research purpose, researchers need to analyze several factors, then find the corresponding data, and use the model to complete. For example, in the authoritative paper *Reconstruction of the full transmission dynamics of COVID-19 in Wuhan*[20] reviewing the Wuhan epidemic, researchers proposed three main factors that must be considered in the development of the COVID-19 epidemic in Wuhan, namely "pre-symptomatic infected individuals", "NPIs" and "human mobility". Subsequently, the researchers used detailed case data and several non-case data to complete the model design with the compartment subdivision, multi-stage model, and dynamic parameters, and finally reviewed the Wuhan epidemic. On this basis, they used the model results to calculate a number of indicators to form valuable conclusions, such as COVID-19's Epidemiological characteristics comparing with SARS and MERS, and judging the effectiveness of NPIs in Wuhan.

4.2 The Relationship between DFM and Research Interests

In the field of epidemic modeling, after being proposed, research would be divided into two parts: model building and simulation experiments. This research process can be summarized in the following steps:

- Research Interest
- Factors
- Origin Data
- Models
- Results of the task

The process of "Factors to Origin Data" and "Origin Data to Models" had been discussed in Chapter 2, and the research methods involved are in Chapter 3. Therefore, the relationship between Factors, Data, and Model(DFM) and the Research Interest will be discussed next, from the aspect of research content.

4.3 Factors and Their DFM

Each type of factor corresponds to some type of data and is associated with several models. The 6 main categories of research factors are summarized below on the basis of reviewing about 100 papers from Science, Nature, PNAS, The Lancet, SIGKDD, TKDE, and AAAI.

4.3.1 Human Mobility

Mobility is a factor related to humans moving from here to there.

In the SEIR model, each individual is uniformly distributed in an ideal space and performs a completely random motion at a uniform speed. After the outbreak of COVID-19, the error brought about by such false assumptions cannot be ignored. Human mobility is not only a factor that directly affects the ability of the epidemic to spread but also acts as a medium for most NPIs, such as isolation, curfews, and school closures, to indirectly affect the spread of the epidemic. Therefore, it is necessary to model crowd mobility factors.

There are two types of data used to model mobility: traffic data with start and end points, and a comprehensive human mobility index. The former is geospatial non-case data, used with the meta-population model, and the latter is array-type non-case data, used with dynamic parameters.

Shengjie Lai's team from Fudan University and others [8] regarded human mobility as a manifestation of NPIs and finally explored the effect of non-pharmaceutical interventions to contain COVID-19 in China by modeling human mobility factors. Using data from mobile phone signaling and travel, such as high-speed rail and plane ticket sales, they calculated the population flow between cities in China. Based on this flow, they completed the modeling of human mobility, which is more in line with the epidemic transmission model in China's NPIs environment than the homogenized SEIR model. Similar data and methods were used by James D. Munday's team from LHSTM (London School of Hygiene and Tropical Medicine) [51]. However, in order to achieve their purpose of studying COVID-19 transmission under school reopening strategies in England, they established a meta-population model with a school-household network structure rather than a normal social network. They used their not-publicly-available data from UK Department for Education (DfE) to construct a network of schools linked through households: each edge on the network of schools is weighted by the number of unique contacts between schools that occur through shared households. For example, if in a given household, 2 children attend school i and 2 children attend school j, this corresponds to 4 unique contacts between school i and school j.

In the comprehensive human mobility index, Google Mobility is a dataset that will be released for free and open to use until the end of COVID-19 [52]. It divides the human activities into 6 categories according to the place of occurrence: Grocery & pharmacy, Parks, Transit stations, Retail & recreation, Residential, and Workplaces, then records the difference between the crowd activity intensity and the reference value at a specific moment in turn. A work from Pierre Nouvellet and his team [53] used comprehensive human mobility indexes such as google mobility to complete the dynamic processing of effective reproduction number at the time of infection ($R_{t,i}$), and finally quantitatively calculated The transmission can be significantly decreased with the initial reduction in mobility in 73%.

In summary, the human-mobility factor is represented by traffic flow data and human mobility index, modeled by meta-population and dynamic parameters.

4.3.2 NPI

Non-pharmaceutical interventions (NPIs) are the factors by which humans, mainly governments or rulers, proactively propose measures to intervene in the epidemic.

Among all six types of factors, NPIs have the most extensive data sources and the most available modeling methods. Like the fundamental position of dynamic parameters in the COVID-19 model, research on COVID-19 must directly or indirectly consider NPIs. Research on indirect NPIs will use other factors, especially human mobility or human response, as a medium for modeling NPIs. Studies that directly consider NPIs use multi-stage models to model NPIs [20].

In section 4.4, factors will be associated with a research interest, where simulation and regression experiments will be reviewed in detail. When dealing with NPIs factors, more researchers consider simulating the effects of NPIs through experimental settings in the simulation experiment part, so as to complete the modeling and result from analysis of NPIs. For example, after using the meta-population model to model other factors such as human mobility, the simulation of the closed isolation policy is realized by manually interfering with the in and out the traffic of a certain population [24].

4.3.3 Ages

Age is a special, highly influential, mainly considered "individual difference within the compartment". In the basic compartment model, individuals in the same compartment are identical, and this inappropriate assumption is addressed by many modeling approaches. Among them, age is more valued by researchers due to two characteristics: compared with other factors such as income and social background, age data has less private information and can be easily counted and utilized; COVID-19 is highly sensitive to age, and there are great differences in the severity rate, mortality, and clinical manifestations of age groups.

The data processing the age factor is Point-feature Data, which contains the proportion of each age composition of

a specific population. Depending on the purpose of the study, the data can be the age ratio of the entire population or the age ratio of a specific diseased/susceptible population. In terms of modeling methods, there are mainly two methods for modeling age factors: unified processing through weighted sum operator; horizontal compartment subdivision according to age.

For example Davies and others[11] published a paper in nature, which systematically analyzed the B.1.1.7 variant outbreak in England from multiple dimensions such as age, region, and medical characteristics. They used the differences in S gene target failures (SGTF) data in PCR tests in different age groups and built a model based on this to complete their research work. Zhang Juanjuan, Yu Hongjie and others[54] sought to study how the COVID-19 outbreak in China is dynamic. They started with the contact pattern, built a detailed age-specific contact coefficient matrix based on age differences, and then built a model to complete the study.

4.3.4 Medical Resources

Medical resources include beds, ICU beds, number of nurses, doctors, ventilators, vaccines, etc. Research on medical resources is based on epidemic prediction, to explore whether medical resources are sufficient and how to allocate them

These research is relatively independent and will add a sub-model dedicated to forecasting medical resource needs on the basis of the compartment model used for epidemic modeling. The data used in forecasting medical resources is divided into two categories: clinical statistics of the Point-feature type, and refined case data of the Array type. Clinical statistics include severe disease rate, mortality rate, ICU utilization rate, average ICU treatment time, etc., which can be a single result or the results by age group and region. On the basis of traditional diagnosis, cure, and death, refined case data adds information such as admission time, onset time, and ICU admission time of these cases.

Institute for Health Metrics and Evaluation (IHME) provides typical research in this area[55]. They first constructed a compartment model for predicting the epidemic and then constructed a data sequence of dead patients of different ages through age data. Using the admission-death time difference data of patients who died in different age groups, they calculated the admission time of patients who died in different age groups, and aligned and summed them on the time axis to obtain the data series of admission times for all patients who died. Based on this, they deduced the total number of hospitalized patients in combination with the death rate of hospitalized patients, and then obtained the predicted results of medical resource demand according to the length of hospitalization of different categories (normal, severe) and the average consumption of various resources.

4.3.5 Human Response

Human response is the fact that people take the initiative to take action out of psychological factors such as fear of death and fear of illness to avoid being infected as much as possible. Such a response will have an impact on the development of an infectious disease based on the principles established by the infectious disease model.

For example, Weitz and others[50] abstract the crowd's fear of death and infection as an awareness factor, then calculate the value of this factor in real-time according to the value of each compartment. They use this factor to correct the parameters of the compartment model. In this way, a self-feedback mechanism based on the dynamic parameter method is built, which models the human response factor.

4.3.6 Vaccine

The vaccine is a broad class of research interests. Vaccine research should first add vaccine-related compartments/parameters to the modeling process, and then obtain results from improved compartment models with vaccines for more detailed and specific epidemiological problems.

A study from Prof. Liu's research team at LHSTM can better represent the data, methods, and models of vaccine research interests. This study[56] examines how COVID-19 epidemic characteristics, population age characteristics, government policies, and population movement all influence optimal vaccine prioritization strategies. In the data phase, in addition to the above general characteristics, the study also introduced data on the distribution rate of 4 vaccines based on COVAX vaccine distribution data and then established the CovidM improved compartment model to obtain the vaccine immunity compartment, vaccine subgroup New bins such as clinical bins are eigenvalues. Finally, these values and statistical experiments were used to obtain the comparison results of the advantages and disadvantages of different vaccine priority distribution strategies on cLE, cQALY, and the other five indicators.

4.3.7 Other Aspect in Society like Economic

The other factors include economic, pollen, UV, and other factors that are not widely studied. There is no universal standard for the data introduced into the study of these factors, or the modeling methods used.

For example, using the results predicted by basic infectious disease model simulations, combined with pollen concentration data, families and others[57] proved through statistical mathematical experiments that higher pollen concentrations are associated with higher rates of COVID-19 transmission. Using a multi-population model, Bonaccorsi and others[58] established a refined infectious disease model based on Italian case data and population flow data, and then used infectious disease indicators and economic indicators to conduct cross-axis statistics to quantitatively calculate the economic changes after the outbreak of COVID-19. Relationship to changes in crowd activity.

4.4 From DFM to Research Interest: Simulation and Regression Experiments

In order to achieve research purposes, researchers use a lot of simulation and regression experiments to process the output results of infectious disease models.

The simulation experiment is based on the infectious disease model itself, by changing some real data, or using hypothetical data as input, and using different model results caused by different inputs to complete the research purpose.

Regression experiments are established outside the infectious disease model, and the output of the infectious disease

model is used as the input to complete the research purpose through mathematical and statistical methods. Since this survey mainly studies the infectious disease model itself, the experimental part is only briefly described.

5 PREPAREDNESS FOR FUTURE OUTBREAKS

The need to predict the timing and intensity of outbreaks of infectious diseases has been acknowledged for quite some time and has emerged as an even stronger lesson from the COVID-19 experience. Current initiatives to address this question depend on strategies for integrating theoretical process models with transmission patterns observed empirically.

Promising research in this area relies on Artificial Intelligence-based tools[59], [60]. The development of a prediction pipeline combines distinct methodologies: machine learning, causal diagrams, and their application to understanding the effect of large-scale drivers (e. g., climate, behavior) on the evolutionary and ecological trajectories leading to the next pandemic. The study of emergent diseases has become a major research topic in biomedicine. Its progress can be attributed to the decisive contributions made possible by the reconstruction of host-pathogen networks as an outcome of machine learning strategies. Interpretable (non-'black box') Machine Learning has proved particularly instrumental in establishing model–lab–field virtuous feedback to challenge and improve predictive models. Cutting-edge success examples focus more on the structural and biochemical interactions between pathogen and cell receptors leading to human–pathogen compatibility. Assessment of the zoonotic potential in this association is the main goal in such studies. Machine Learning approaches that can predict protein folding structures are expected to be added to the current toolbox already containing pipelines to harness the power of the whole individual and population genomes. Actionable predictions require robustness and interpretability of outputs generated by pipelines based on Machine Learning strategies. This might be achievable by embedding AI algorithms with the capability to find causes[61]. Artificial Intelligence and Causal Inference are research areas that experienced formidable progress in the last decades but only now are finding a common ground.

Such framework has been applied to assess the causal role of environmental changes as drivers of vector-borne diseases[62]. Santos et al. examined this relationship in detail using the spread of visceral leishmaniasis (VL) in São Paulo state (Brazil) as the case study. A two-step approach estimated the causal effects (overall, direct, and indirect) of deforestation on the occurrence of the VL vector, canine visceral leishmaniasis (CVL), and human visceral leishmaniasis (HVL).

Integration of Machine Learning and Causal Inference approaches raises the expectations of researchers in this field and might provide the appropriate methodological tools to face the challenges we encounter when studying the effect of large-scale drivers (e. g., climate) on human diseases. To achieve this, we need to conceive causal diagrams addressing evolutionary and ecological processes, an area yet almost untouched[63]. Such diagrams provide the necessary framework to describe the causal effects of events that take place at distinct interacting levels, such as ecological (deforestation), evolutionary (founder effect, niche construction), and genetic (genetic basis of vector-pathogen competence), thus helping us in forecasting the future host-pathogen landscape under climate change and intervention strategies.

6 SEARCH STRATEGIES

6.1 Search Words

COVID-19 modeling, Sars-cov-2, Compartment model, Agent-based model, AI infectious disease modeling, Vaccine modeling, infectious disease prediction and simulation.

6.2 Search Resource

- **Journals and Conferences**:Nature, Science, PNAS, KDD, AAAI, TKDE, The lancet, NEJM, JAMA.
- **Databases**:Google Mobility, Oxford Government Response Index, IHME medical resource.
- **websites**:Github, Google, Readthedocs(docs for covasim/openABM, etc.).

6.3 Inclusion and Exclusion Criteria

- **Date**:researches after 2020.01.01,
- **Exposure of interest**:infectious disease modeling,
- **Geographic location of study**:None,
- **Language**:English,
- **Participants**:at least 1 professor,
- **Peer review**:None,
- **Reported outcomes**:None,
- **Setting**:None,
- **Study design**:build a model to simulate/predict/analysis COVID-19,
- **Type of Publication**:Nature/Science, PNAS, Top Journals/Conferences in Epidemiology or Computer Science.

7 CONCLUSION

After the outbreak of COVID-19, traditional compartment models, like SEIR, which is commonly used to model infectious diseases encountered extreme limitations. The main source of this limitation is the various influences that the social system exerts on the infectious disease system. To analyze these effects, researchers summarize them into abstract factors, then find the corresponding data representation factors, apply appropriate modeling methods to the data, and complete the study through experiments.

This review does not divide the collection of papers according to the research purpose, but starts from the process of infectious disease modeling research, corresponds to the data, methods, models and research interests, and conducts survey work on the cross-sectional area of the workflow. In terms of modeling methods, the researchers use compartment subdivision, dynamic parameters, agent-based model methods and AI-related methods. The compartment subdivision and dynamic parameters in turn optimize the structure and value of the compartment model, and the agent model changes the compartment model's

assumptions about the "compartment". AI-assisted or AI-led models offer a new avenue for modeling infectious diseases.

In terms of factors studied, the researchers studied 6 categories: human mobility, NPIs, ages, medical resources, human response, and vaccine. Using data or modeling methods, researchers try to achieve the modeling effect that is closest to the actual situation, so as to provide their own suggestions for quantitative analysis of the impact of social systems, and even for the future transmission status of infectious diseases and prevention and control strategies.

Appendix A
Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution

Honghao Shi completed the review research, chapter design and article writing, Jingyuan Wang provided guidance and revisions for all sections, Jiawei Cheng assisted with the research and writing of Chapter 3 on compartment segmentation and dynamic models, Xiaopeng Qi and Hanran Ji assisted the specific application and business of Chapter 4, Claudio J. Struchiner and Daniel A.M. Villela completed the work on future epidemic prediction, supplementing the application scenarios of this article, Eduard V. Karamov and Ali S. Turgiev completed the design, application and research of the agent-based model.

Funding

We received project support and design guidance from National Key R&D Program of China (2021ZD0111201), The National Natural Science Foundation of China (Grant No. 82161148011, 72171013), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – Refs. 441057/2020-9, 309569/2019-2), CJS - CNPq, and Fundação deAmparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ), and The Russian Foundation for basic Research, project number 21-51-80000.

References

[1] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[2] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[3] Matt J Keeling and Bryan T Grenfell. Understanding the persistence of measles: reconciling theory, simulation and observation. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1489):335–343, 2002.

[4] Marc Lipsitch, Ted Cohen, Ben Cooper, and et al. Transmission dynamics and control of severe acute respiratory syndrome. *science*, 300(5627):1966–1970, 2003.

[5] Duygu Balcan, Hao Hu, Bruno Goncalves, and et al. Seasonal transmission potential and activity peaks of the new influenza a (h1n1): a monte carlo likelihood analysis based on human mobility. *BMC medicine*, 7(1):1–12, 2009.

[6] Maureen Hurley, Glen Jacobs, and Melinda Gilbert. The basic si model. *New Directions for Teaching and Learning*, 2006(106):11–22, 2006.

[7] Osmar Pinto Neto, Deanna M Kennedy, José Clark Reis, and et al. Mathematical model of covid-19 intervention scenarios for são paulo—brazil. *Nature communications*, 12(1):1–13, 2021.

[8] Shengjie Lai, Nick W Ruktanonchai, Liangcai Zhou, and et al. Effect of non-pharmaceutical interventions to contain covid-19 in china. *nature*, 585(7825):410–413, 2020.

[9] Darlan S Candido, Ingra M Claro, Jaqueline G De Jesus, and et al. Evolution and epidemic spread of sars-cov-2 in brazil. *Science*, 369(6508):1255–1260, 2020.

[10] Huaiyu Tian, Yonghong Liu, Yidan Li, and et al. An investigation of transmission control measures during the first 50 days of the covid-19 epidemic in china. *Science*, 368(6491):638–642, 2020.

[11] Nicholas G Davies, Sam Abbott, Rosanna C Barnard, and et al. Estimated transmissibility and impact of sars-cov-2 lineage b. 1.1. 7 in england. *Science*, 372(6538):eabg3055, 2021.

[12] Erik Volz, Swapnil Mishra, Meera Chand, and et al. Assessing transmissibility of sars-cov-2 lineage b. 1.1. 7 in england. *Nature*, 593(7858):266–269, 2021.

[13] Stephen M Kissler, Christine Tedijanto, Edward Goldstein, and et al. Projecting the transmission dynamics of sars-cov-2 through the postpandemic period. *Science*, 368(6493):860–868, 2020.

[14] Ruiyun Li, Bin Chen, Tao Zhang, and et al. Global covid-19 pandemic demands joint interventions for the suppression of future waves. *Proceedings of the National Academy of Sciences*, 117(42):26151–26157, 2020.

[15] Tamma Carleton, Jules Cornetet, Peter Huybers, and et al. Global evidence for ultraviolet radiation decreasing covid-19 growth rates. *Proceedings of the National Academy of Sciences*, 118(1), 2021.

[16] Rachel E Baker, Wenchang Yang, Gabriel A Vecchi, and et al. Susceptible supply limits the role of climate in the early sars-cov-2 pandemic. *Science*, 369(6501):315–319, 2020.

[17] Ling Xue, Shuanglin Jing, Joel C Miller, and et al. A data-driven network model for the emerging covid-19 epidemics in wuhan, toronto and italy. *Mathematical Biosciences*, 326:108391, 2020.

[18] Henrik Salje, Cécile Tran Kiem, Noémie Lefrancq, and et al. Estimating the burden of sars-cov-2 in france. *Science*, 369(6500):208–211, 2020.

[19] Vadim A Karatayev, Madhur Anand, and Chris T Bauch. Local lockdowns outperform global lockdown on the far side of the covid-19 epidemic curve. *Proceedings of the National Academy of Sciences*, 117(39):24575–24580, 2020.

[20] Xingjie Hao, Shanshan Cheng, Degang Wu, and et al. Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature*, 584(7821):420–424, 2020.

[21] Daniel Duque, David P Morton, Bismark Singh, and et al. Timing social distancing to avert unmanageable covid-19 hospital surges. *Proceedings of the National Academy of Sciences*, 117(33):19873–19878, 2020.

[22] Benjamin F Maier and Dirk Brockmann. Effective containment explains subexponential growth in recent confirmed covid-19 cases in china. *Science*, 368(6492):742–746, 2020.

[23] Thomas Hale, Noam Angrist, Rafael Goldszmidt, and et al. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature Human Behaviour*, 5(4):529–538, 2021.

[24] Serina Chang, Emma Pierson, Pang Wei Koh, and et al. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

[25] Sheryl L Chang, Nathan Harding, Cameron Zachreson, and et al. Modelling transmission and control of the covid-19 pandemic in australia. *Nature communications*, 11(1):1–13, 2020.

[26] Amray Schwabe, Joel Persson, and Stefan Feuerriegel. Predicting covid-19 spread from large-scale mobility data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3531–3539, 2021.

[27] Alessio Andronico, Cécile Tran Kiem, Juliette Paireau, and et al. Evaluating the impact of curfews and other measures on sars-cov-2 transmission in french guiana. *Nature communications*, 12(1):1–8, 2021.

[28] Abiel Sebhatu, Karl Wennberg, Stefan Arora-Jonsson, and et al. Explaining the homogeneous diffusion of covid-19 nonpharmaceutical interventions across heterogeneous countries. *Proceedings of the National Academy of Sciences*, 117(35):21201–21208, 2020.

[29] Tobias S Brett and Pejman Rohani. Transmission dynamics reveal the impracticality of covid-19 herd immunity strategies. *Proceedings of the National Academy of Sciences*, 117(41):25897–25903, 2020.

[30] Valentina Marziano, Giorgio Guzzetta, Bruna Maria Rondinone, and et al. Retrospective analysis of the italian exit strategy from covid-19 lockdown. *Proceedings of the National Academy of Sciences*, 118(4), 2021.

[31] Tom Britton, Frank Ball, and Pieter Trapman. A mathematical model reveals the influence of population heterogeneity on herd immunity to sars-cov-2. *Science*, 369(6505):846–849, 2020.

[32] Laura Di Domenico, Giulia Pullano, Chiara E Sabbatini, and et al. Modelling safe protocols for reopening schools during the covid-19 pandemic in france. *Nature communications*, 12(1):1–10, 2021.

[33] Stefan Thurner, Peter Klimek, and Rudolf Hanel. A network-based explanation of why most covid-19 infection curves are linear. *Proceedings of the National Academy of Sciences*, 117(37):22684–22689, 2020.

[34] Frank Schlosser, Benjamin F Maier, Olivia Jack, and et al. Covid-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proceedings of the National Academy of Sciences*, 117(52):32883–32890, 2020.

[35] Parham Azimi, Zahra Keshavarz, Jose Guillermo Cedeno Laurent, and et al. Mechanistic transmission modeling of covid-19 on the diamond princess cruise ship demonstrates the importance of aerosol transmission. *Proceedings of the National Academy of Sciences*, 118(8), 2021.

[36] Felix Wong and James J Collins. Evidence that coronavirus superspreading is fat-tailed. *Proceedings of the National Academy of Sciences*, 117(47):29416–29418, 2020.

[37] Nicholas Kortessis, Margaret W Simon, Michael Barfield, and et al. The interplay of movement and spatiotemporal variation in transmission degrades pandemic control. *Proceedings of the National Academy of Sciences*, 117(48):30104–30106, 2020.

[38] Sheikh Taslim Ali, Lin Wang, Eric HY Lau, and et al. Serial interval of sars-cov-2 was shortened over time by nonpharmaceutical interventions. *Science*, 369(6507):1106–1109, 2020.

[39] Akihiro Nishi, George Dewey, Akira Endo, and et al. Network interventions for managing the covid-19 pandemic and sustaining economy. *Proceedings of the National Academy of Sciences*, 117(48):30285–30294, 2020.

[40] Max SY Lau, Bryan Grenfell, Michael Thomas, and et al. Characterizing superspreading events and age-specific infectiousness of sars-cov-2 transmission in georgia, usa. *Proceedings of the National Academy of Sciences*, 117(36):22430–22435, 2020.

[41] Kim Sneppen, Bjarke Frost Nielsen, Robert J Taylor, and et al. Overdispersion in covid-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. *Proceedings of the National Academy of Sciences*, 118(14), 2021.

[42] Jesús A Moreno López, Beatriz Arregui García, Piotr Bentkowski, and et al. Anatomy of digital contact tracing: Role of age, transmission setting, adoption, and case detection. *Science advances*, 7(15):eabd8750, 2021.

[43] 1. G. N. Rykovanov, S. N. Lebedev, and et al O. V. Zatsepin. Agent-based simulation of the covid-19 epidemic in russia. *Herald of the Russian Academy of Sciences*, 92(4):479–487, 2021.

[44] Cliff C Kerr, Robyn M Stuart, Dina Mistry, and et al. Covasim: an agent-based model of covid-19 dynamics and interventions. 17(7):e1009149, 2021.

[45] Robert Hinch, William JM Probert, Anel Nurtay, and et al. Openabm-covid19—an agent-based model for non-pharmaceutical interventions against covid-19 including contact tracing. *PLoS computational biology*, 17(7):e1009146, 2021.

[46] Jared Willard, Xiaowei Jia, Shaoming Xu, and et al. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys (CSUR)*, 2021.

[47] Salah Ghamizi, Renaud Rwemalika, Maxime Cordy, and et al. Data-driven simulation and optimization for covid-19 exit strategies. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3434–3442, 2020.

[48] Lijing Wang, Aniruddha Adiga, Jiangzhuo Chen, and et al. Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. 2022.

[49] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[50] Joshua S Weitz, Sang Woo Park, Ceyhun Eksin, and et al. Awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations. *Proceedings of the National Academy of Sciences*, 117(51):32764–32771, 2020.

[51] James D Munday, Katharine Sherratt, Sophie Meakin, and et al. Implications of the school-household network structure on sars-cov-2 transmission under school reopening strategies in england. *Nature communications*, 12(1):1–11, 2021.

[52] Google LLC. Google covid-19 community mobility reports. [EB/OL]. https://www.google.com/covid19/mobility/ Accessed: ¡2022-02-18¿.

[53] Pierre Nouvellet, Sangeeta Bhatia, Anne Cori, and et al. Reduction in mobility and covid-19 transmission. *Nature communications*, 12(1):1–9, 2021.

[54] Juanjuan Zhang, Maria Litvinova, Yuxia Liang, and et al. Changes in contact patterns shape the dynamics of the covid-19 outbreak in china. *Science*, 368(6498):1481–1486, 2020.

[55] IHME COVID, Christopher JL Murray, et al. Forecasting the impact of the first wave of the covid-19 pandemic on hospital demand and deaths for the usa and european economic area countries. *MedRxiv*, 2020.

[56] Yang Liu, Frank G Sandmann, Rosanna C Barnard, and et al. Optimising health and economic impacts of covid-19 vaccine prioritisation strategies in the who european region: a mathematical modelling study. *The Lancet Regional Health-Europe*, 12:100267, 2022.

[57] Athanasios Damialis, Stefanie Gilles, Mikhail Sofiev, and et al. Higher airborne pollen concentrations correlated with increased sars-cov-2 infection rates, as evidenced from 31 countries across the globe. *Proceedings of the National Academy of Sciences*, 118(12):e2019034118, 2021.

[58] Giovanni Bonaccorsi, Francesco Pierri, Matteo Cinelli, and et al. Economic and social consequences of human mobility restrictions under covid-19. *Proceedings of the National Academy of Sciences*, 117(27):15530–15535, 2020.

[59] Gregory F Albery, Daniel J Becker, Liam Brierley, and et al. The science of the host–virus network. *Nature microbiology*, 6(12):1483–1492, 2021.

[60] Cheng Zhang and Frederick A Matsen IV. Generalizing tree probability estimation via bayesian networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[61] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[62] Cleber Vinicius Brito dos Santos, Anaia da Paixao Seva, Guilherme Loureiro Werneck, and et al. Does deforestation drive visceral leishmaniasis transmission? a causal analysis. *Proceedings of the Royal Society B*, 288(1957):20211537, 2021.

[63] Jun Otsuka. Causal foundations of evolutionary genetics. *The British Journal for the Philosophy of Science*, 2020.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: