Innov Biosyst Bioeng, 2020, vol. 4, no. 1, 13–18
doi: 10.20535/ibb.2020.4.1.195074
UDC 519.2, 519.8

Original article

STATISTICS-BASED PREDICTIONS OF CORONAVIRUS EPIDEMIC SPREADING IN MAINLAND CHINA

I. Nesteruk*

Institute of Hydromechanics, National Academy of Sciences of Ukraine, Kyiv, Ukraine
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine

*Corresponding author: inesteruk@yahoo.com

Received 10 February 2020; Accepted 18 February 2020

Background. The epidemic outbreak caused by coronavirus COVID-19 is of great interest to researches because of the high rate of the infection spread and the significant number of fatalities. A detailed scientific analysis of the phenomenon is yet to come, but the public is already interested in the questions of the epidemic duration, the expected number of patients and deaths. Long-time predictions require complicated mathematical models that need a lot of effort to identify and calculate unknown parameters. This article will present some preliminary estimates.

Objective. Since the long-time data are available only for mainland China, we will try to predict the epidemic characteristics only in this area. We will estimate some of the epidemic characteristics and present the dependencies for victim numbers, infected and removed persons versus time.

Methods. In this study we use the known SIR model for the dynamics of an epidemic, the known exact solution of the linear differential equations and statistical approach developed before for investigation of the children disease, which occurred in Chernivtsi (Ukraine) in 1988–1989.

Results. The optimal values of the SIR model parameters were identified with the use of statistical approach. The numbers of infected, susceptible and removed persons versus time were predicted and compared with the new data obtained after February 10, 2020, when the calculations were completed.

Conclusions. The simple mathematical model was used to predict the characteristics of the epidemic caused by coronavirus in mainland China. Unfortunately, the number of coronavirus victims is expected to be much higher than that predicted on February 10, 2020, since 12289 new cases (not previously included in official counts) have been added two days later. Further research should focus on updating the predictions with the use of up-to-date data and using more complicated mathematical models.

Keywords: coronavirus epidemic in China; coronavirus COVID-19; coronavirus 2019-nCoV; mathematical modeling of infection diseases; SIR model; parameter identification; statistical methods.

Introduction

Here, we consider the development of an epidemic outbreak caused by coronavirus COVID-19 (the previous name was 2019-nCoV) (see e.g., [1–3]). Since long-term data are available only for mainland China, we will try to predict the number of coronavirus victims V (number of persons who caught the infection and got sick) only in this area. The first estimations of $V(t)$ exponential growth versus time t, typical for the initial stages of every epidemic (see e.g., [4]) have been done in [3]. For long-time predictions, more complicated mathematical models are necessary. For example, a susceptible-exposed-infectious-recovered (SEIR) model was used in [2]. Nevertheless, complicated models need more effort for unknown parameters identification. This procedure may be especially difficult if reliable data are limited.

In this study, we use the known SIR model for the dynamics of an epidemic [4–8]. For the parameter identification, we will use the exact solution of the SIR set of linear equations and statistical approach developed in [4] (tested also in [9]). These methods were applied for investigation of the children disease, which occurred in Chernivtsi (Ukraine) in 1988–1989. We will estimate some of the epidemic characteristics and present the dependencies for victim numbers, infected and removed persons versus time.

Materials and Methods

Data

We shall analyze the daily data for the number of confirmed cases in mainland China, which origins from the National Health Commission of the People's Republic of China [1]. A part of the official

diagram (its version, presented on February 15, 2020) is shown in Fig. 1. For calculations, we have used the data for the period of time from January 16 to February 9, 2020. The numbers shown after February 9 were used for verification of predictions.

On February 12, 2020, the National Health Commission of the People's Republic of China has added 12289 new cases (not previously included in official counts) as "clinically diagnosed cases". The cases, reported by this official organization before, have the name of "tested confirmed cases" [1]. To avoid confusiong, we will denote "tested confirmed cases" as W_j; j corresponds to the different time moments t_j (see the Table). Let us denote the "clinically diagnosed cases" as Q_j. The sum of W_j and Q_j is shown in the last column in Fig. 1 and in the Table.

The Table shows that the precise time of the epidemic beginning t_0 is unknown. Therefore, the optimization procedures have to determine the optimal value of this parameter as well as for other parameters of SIR model.

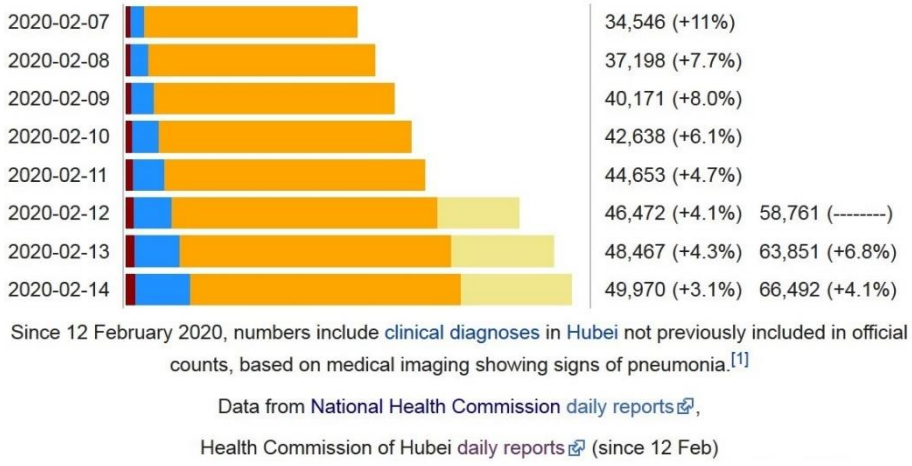

Figure 1: A part of official diagram with the numbers of $W_j + Q_j$ (last column) and W_j (previous column) [1]

Table: The information from the official table of the National Health Commission of the People's Republic of China [1]. The corresponding time moments t_j and the number of W_j and Q_j which were used for calculations and verification

Day in January, 2020	Time moment t_j	"Tested confirmed cases" W_j	Day in February, 2020	Time moment t_j	"Tested confirmed cases" W_j	The sum of "tested confirmed cases" and "clinically diagnosed cases" $W_j + Q_j$
16	0	45	1	16	14380	Unknown
17	1	62	2	17	17205	Unknown
18	2	121	3	18	20440	Unknown
19	3	198	4	19	24324	Unknown
20	4	291	5	20	28018	Unknown
21	5	440	6	21	31161	Unknown
22	6	571	7	22	34568	Unknown
23	7	830	8	23	37198	Unknown
24	8	1287	9	24	40171	Unknown
25	9	1975	10	25	42638	Unknown
26	10	2744	11	26	44653	Unknown
27	11	4515	12	27	46472	58761
28	12	5974	13	28	48467	63851
29	13	7711	14	29	49970	66492
30	14	9692	–	–	–	–
31	15	11791	–	–	–	–

Exact solution of SIR equations

The SIR model for an infectious disease can be written as follows [6, 7]:

$$\dot{S} = -\alpha SI, \quad (1)$$

$$\dot{I} = \alpha SI - \rho I, \quad (2)$$

$$\dot{R} = \rho I. \quad (3)$$

The number of susceptible persons is S, infected (persons who are sick and spread the infection) – I, removed (persons who do not spread the infection anymore, this number is the sum of isolated, recovered and dead people) – R; the infection and immunization rates are α and ρ respectively. Since $\dot{S} + \dot{I} + \dot{R} = 0$ (see, eqs. (1)–(3)), the sum $N = S + I + R$ must be constant for all moments of time and can be treated as the amount on susceptible persons before the outbreak of an epidemic, since $I = R = 0$ at $t < t_0$. It must be noted that the constant N is not the volume of population N_{total} but only the initial number of people sensitive and not protected to some specific disease. In particular, the ratio N/N_{total} may be rather small.

To determine the initial conditions for the set of eqs. (1)–(3), let us suppose that

$$I(t_0) = 1, \quad R(t_0) = 0, \quad S(t_0) = N - 1. \quad (4)$$

It follows from (1) and (2) that

$$\frac{dI}{dS} = \frac{\nu}{S} - 1, \quad \nu = \frac{\rho}{\alpha}. \quad (5)$$

Integration of (5) with the initial conditions (4) yields:

$$I = \nu \ln S - S + N - \nu \ln(N - 1). \quad (6)$$

Function I has a maximum at $S = \nu$ and tends to zero at infinity, see [6, 7]. In comparison, the number of susceptible persons at infinity $S_\infty > 0$, and can be calculated with the use of (6) from a non-linear equation

$$S_\infty = (N - 1)e^{\frac{S_\infty - N}{\nu}}. \quad (7)$$

In [4] the equations (1)–(3) were solved by introducing the function $V(t) = I(t) + R(t)$, corresponding to the number of victims. The integration of corresponding equation

$$\dot{V} = \alpha SI = \alpha(N - V) \times [\nu \ln(N - V) + V - \nu \ln(N - 1)] \quad (8)$$

yields:

$$t = \frac{F_1(V, N, \nu) + \alpha t_0}{\alpha}, \quad (9)$$

$$F_1 = \int_1^V \frac{dU}{(N - U)[\nu \ln(N - U) + U - \nu \ln(N - 1)]}. \quad (10)$$

Thus, for every set of parameters N, ν, α, t_0 and a fixed value of V the integral (10) can be calculated and the corresponding moment of time can be determined from (9). Then I can be calculated from (6) by putting $S = N - V$ and function R from $R = V - I$.

Statistical approach for parameter identification. Linear regression

As in paper [4], we shall use the fact that the random function $F_1(V, N, \nu)$ has a linear distribution (see (9)). Then we can apply the linear regression (see [10]) for every pair of parameters N and ν and calculate the corresponding values of t_0 and α. The optimal (the most reliable) values of N and ν correspond to the maximum value of the correlation coefficient r (see [4, 9]).

Results

Since we did not know and still don't know the values of Q_j before February 12, 2020, we supposed that $V_j = W_j$ and have done the calculations with the use of data for the time period from January 16 to February 9, 2020. The optimal values of the parameters are:

$$N = 90611;$$

$$\nu = 65546.5;$$

$$\alpha = 1.477985357571669e - 05;$$

$$t_0 = -7.720998173432072.$$

The corresponding correlation coefficient is very high $r = 0.997966487046645$. The solution of (7) yields the value $S_\infty = 45579$. The corresponding number of infected I, susceptible S and removed R persons versus time (starting from January 16, 2020) were calculated and shown in Fig. 2. The blue line represents the number of victims $V = I + R$ and is in good agreement with "tested confirmed cases" W_j, reported by the National Health Commission of the People's Republic of China [1] (blue markers).

Discussion

Unfortunately, many cases have not been included in the official counts and have appeared in the official Table from [1] only on February 12 as "clinically diagnosed cases" Q_j (see Fig. 1). Since the National Health Commission of the the People's Republic of China has proposed two different ways of registration of the same disease [1], V_j must be the sum of W_j and Q_j, i.e. $V_j = W_j + Q_j$ (provided that no new methods of registering the same disease would appear). Values W_j after February 9 are shown in Fig. 3 by "stars". "Crosses" represent the sum $W_j + Q_j$.

Since the optimal curve was obtained only with the use of W_j and the difference between W_j and V_j is very big (e.g., it was 12 289 persons on February 12, 2020), the predictions shown in Fig. 2 and reported in [11] are no longer relevant. To have better predictions, it is necessary to have exact Q_j – data for the period before February 12.

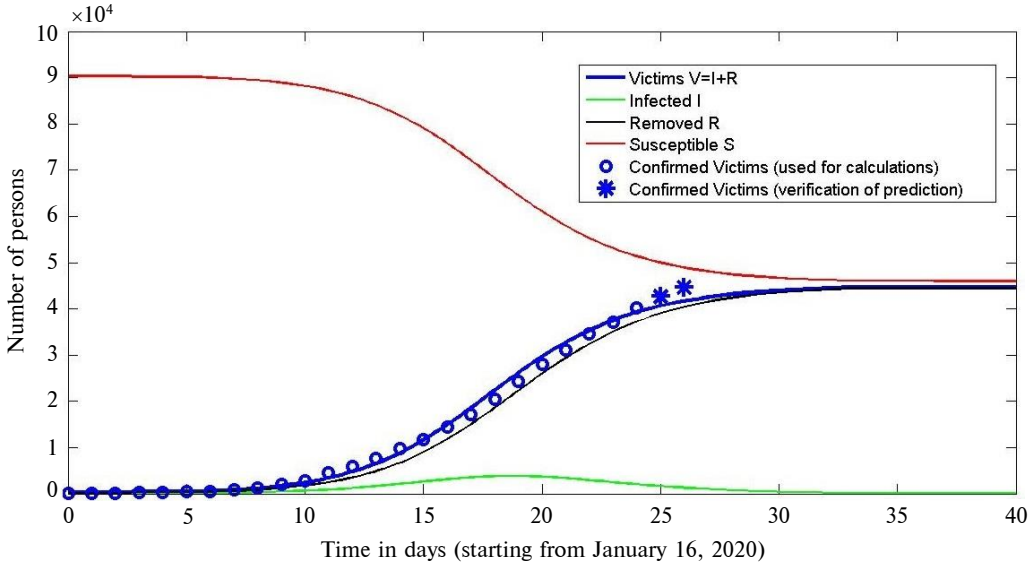

Figure 2: Results of calculations and verification

Numbers of infected I (green line), susceptible S (red line), and removed R (black line) persons versus time in days (starting from January 16, 2020). The blue line represents the number of victims $V = I + R$. Blue markers show the "tested confirmed cases" W_j, reported by the National Health Commission of the People's Republic of China [1]. The "circles" correspond to the points used for calculations (it was supposed that $V_j = W_j$); "stars" – to the points used only for verification

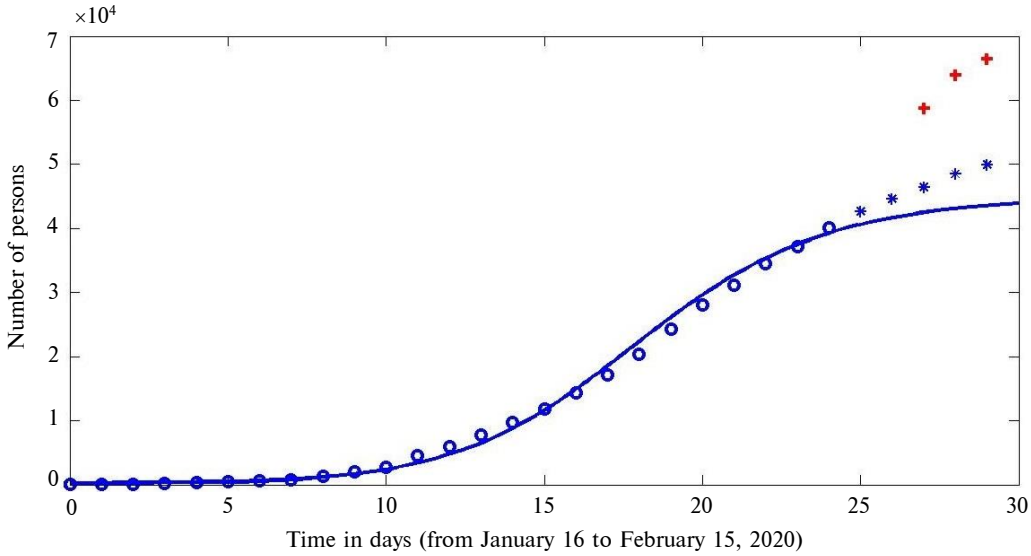

Figure 3: Prediction (line) and observations from [1] (markers)

"Circles" show the "tested confirmed cases" W_j for the period from January 16 to February 9, 2020, [2]. These points were used to calculate the prediction curve. "Stars" correspond to the "tested confirmed cases" W_j for the period from February 10 to February 14, 2020, [1]. "Crosses" represent the sum $W_j + Q_j$ from [1]

Conclusions

The simple mathematical model was used to predict the characteristics of the epidemic caused by coronavirus in mainland China. The numbers of infected, susceptible, and removed persons versus time were predicted and compared with the new data obtained after February 10, 2020, when the calculations were completed. Unfortunately, many cases have not been included in the official counts and have appeared on February 12 only. It makes the predictions reported on February 10, 2020, no longer relevant. Further research should focus on updating the predictions with the use of corrected data and more complicated mathematical models.

Acknowledgements

I would like to express my sincere thanks to professors Dirk Langemann (Techniche Universität Braunschweig) and Juergen Prestin (Universität zu Lübeck) for their support in developing the used optimization approach. I would like to thank also professors Alberto Redaelli, Giuseppe Passoni and Gianfranco Fiore (Politecnico di Milano), Sergei Pereverzyev (RICAM, Linz, Austria) for involving me in very interesting biomedical investigations in frames of EU-financed Horizon-2020 projects EUMLS (Grant agreement PIRSES-GA-2011-295164-EUMLS) and AMMODIT (Grant Number MSCA-RISE 645672).

References

[1] Timeline of the 2019−20 Wuhan coronavirus outbreak [Internet]. En.wikipedia.org. 2020 [cited 2020 Feb 15]. Available from: https://en.wikipedia.org/wiki/Timeline_of_the_2019%E2%80%9320_Wuhan_coronavirus_outbreak

[2] Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. Lancet. 2020 Jan 31. DOI: 10.1016/S0140-6736(20)30260-9

[3] Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. Int J Infect Dis. 2020 Jan 30. DOI: 10.1016/j.ijid.2020.01.050

[4] Nesteruk I. Statistics based models for the dynamics of Chernivtsi children disease. Naukovi Visti NTUU KPI. 2017;5:26-34. DOI: 10.20535/1810-0546.2017.5.108577

[5] Kermack WD, McKendrick AG. A Contribution to the mathematical theory of epidemics. J Royal Stat Soc Ser A. 1927;115:700-21.

[6] Murray JD. Mathematical Biology I/II. New York: Springer; 2002.

[7] Bailey NTJ. The mathematical theory of epidemics. Griffin Book Co.; 1957.

[8] Langemann D, Nesteruk I, Prestin J. Comparison of mathematical models for the dynamics of the Chernivtsi children disease. Mathematics in Computers and Simulation. 2016;123:68-79. DOI: 10.1016/j.matcom.2016.01.003

[9] Nesteruk I. Maximal speed of underwater locomotion. Innov Biosyst Bioeng. 2019;3(3):152-67. DOI: 10.20535/ibb.2019.3.3.177976

[10] Draper NR, Smith H. Applied regression analysis. 3rd ed. John Wiley; 1998.

[11] Nesteruk I. Statistics based predictions of coronavirus 2019-nCoV spreading in mainland China. MedRxiv. 2020 Feb 13. DOI: 10.1101/2020.02.12.20021931

І.Г. Нестерук

ПРОГНОЗИ ПОШИРЕННЯ ЕПІДЕМІЇ КОРОНАВІРУСУ В МАТЕРИКОВОМУ КИТАЇ НА ОСНОВІ СТАТИСТИЧНИХ ДАНИХ

Проблематика. Епідемія, спричинена коронавірусом COVID-19, становить великий інтерес для дослідників через високу швидкість поширення інфекції та значну кількість смертельних випадків. Детальний науковий аналіз цього явища ще попереду, але громадськість уже зацікавлена питаннями тривалості епідемії, очікуваної кількості хворих та смертності. Для довгострокових прогнозів необхідні складні математичні моделі, які потребують багатьох зусиль для ідентифікації невідомих параметрів та обчислень. У цій статті будуть представлені деякі попередні оцінки.

Мета. Оскільки дані за достатньо довгий період часу доступні лише для материкового Китаю, ми спробуємо передбачити характеристики епідемії лише в цьому регіоні. Ми оцінимо деякі характеристики епідемії та подамо залежності від часу кількості жертв, інфікованих та вилучених осіб.

Методика реалізації. У цьому дослідженні ми використовуємо відому SIR-модель для динаміки епідемії, відомий точний розв'язок системи лінійних диференціальних рівнянь і статистичний підхід, розроблений раніше для дослідження дитячої хвороби, що сталася в Чернівцях (Україна) у 1988–1989 рр.

Результати. Оптимальні значення параметрів SIR-моделі були визначені за допомогою статистичного підходу. Кількість заражених, сприйнятливих та вилучених осіб залежно від часу прогнозувалась та порівнювалась із новими даними, отриманими після 10 лютого 2020 р., коли розрахунки були завершені.

Висновки. Для прогнозування особливостей епідемії, спричиненої коронавірусом у материковому Китаї, використовувалась проста математична модель. На жаль, очікується, що кількість жертв коронавірусу буде значно більшою, ніж прогнозувалося 10 лютого 2020 р., оскільки через два дні було додано 12289 нових випадків (раніше не включених до офіційних підрахунків).

Подальші дослідження варто зосередити на оновленні прогнозів на основі свіжих даних та з використанням більш складних математичних моделей.

Ключові слова: епідемія коронавірусу в Китаї; коронавірус COVID-19; коронавірус 2019-nCoV; математичне моделювання інфекційних захворювань; SIR-модель; ідентифікація параметрів; статистичні методи.

И.Г. Нестерук

ПРОГНОЗЫ РАСПРОСТРАНЕНИЯ ЭПИДЕМИИ КОРОНАВИРУСА В МАТЕРИКОВОМ КИТАЕ НА ОСНОВЕ СТАТИСТИЧЕСКИХ ДАННЫХ

Проблематика. Эпидемическая вспышка, вызванная коронавирусом COVID-19, представляет большой интерес для исследователей из-за высокой скорости распространения инфекции и значительного числа умерших. Подробный научный анализ этого явления еще впереди, но общественность уже интересуется вопросами продолжительности эпидемии, ожидаемого числа пациентов и случаев смерти. Для долгосрочных прогнозов необходимы сложные математические модели, которые требуют много усилий для идентификации неизвестных параметров и расчетов. В этой статье будут представлены некоторые предварительные оценки.
Цель. Поскольку данные за долгое время доступны только для материкового Китая, мы попытаемся предсказать характеристики эпидемии только в этом регионе. Мы оценим некоторые из характеристик эпидемии и представим зависимости от времени числа пострадавших, инфицированных и удаленных людей.
Методика реализации. В этом исследовании мы используем известную SIR-модель для динамики эпидемии, известное точное решение системы линейных дифференциальных уравнений и статистический подход, разработанный ранее для исследования детской болезни, которая случилась в Черновцах (Украина) в 1988–1989 гг.
Результаты. Оптимальные значения параметров SIR-модели были определены с использованием статистического подхода. Число инфицированных, восприимчивых и удаленных людей в зависимости от времени было предсказано и сопоставлено с новыми данными, полученными после 10 февраля 2020 г., когда расчеты были завершены.
Выводы. Для прогнозирования характеристик эпидемии, вызванной коронавирусом в материковом Китае, использовалась простая математическая модель. К сожалению, ожидается, что число жертв коронавируса в материковом Китае будет намного выше, чем прогнозировалось 10 февраля 2020 г., поскольку через два дня было добавлено 12289 новых случаев (ранее не включенных в официальные подсчеты). Дальнейшие исследования должны быть направлены на обновление прогнозов на основе свежих данных и с использованием более сложных математических моделей.

Ключевые слова: эпидемия коронавируса в Китае; коронавирус COVID-19; коронавирус 2019-nCoV; математическое моделирование инфекционных заболеваний; SIR-модель; идентификация параметров; статистические методы.