A Kermack–McKendrick model with age of infection starting from a single or multiple cohorts of infected patients

Jacques Demongeot, Quentin Griette, Yvon Maday, Pierre Magal

PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Demongeot J, Griette Q, Maday Y, Magal P. 2023 A Kermack–McKendrick model with age of infection starting from a single or multiple cohorts of infected patients. *Proc. R. Soc. A* **479**: 20220381.
https://doi.org/10.1098/rspa.2022.0381

Received: 17 June 2022
Accepted: 6 January 2023

Subject Areas:
integral equations, differential equations

Keywords:
age of infection, epidemic model, single and multiple cohorts, cluster of infected patients, daily reproduction number, Volterra integral equations

Author for correspondence:
Pierre Magal
e-mail: pierre.magal@u-bordeaux.fr

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6500603.

THE ROYAL SOCIETY PUBLISHING

A Kermack–McKendrick model with age of infection starting from a single or multiple cohorts of infected patients

Jacques Demongeot[1], Quentin Griette[2], Yvon Maday[3,4] and Pierre Magal[5,6]

[1]Université Grenoble Alpes, AGEIS EA7407, 38700, La Tronche, France
[2]UNIHAVRE, LMAH, FR-CNRS-3335, ISCN, Normandie Univ, 76600, Le Havre, France
[3]Sorbonne Université, CNRS, Université Paris Cité, Laboratoire Jacques-Louis Lions (LJLL), F-75005 Paris, France
[4]Institut Universitaire de France, 75005 Paris, France
[5]Univ. Bordeaux, IMB, UMR, 5251, 33400 Talence, France
[6]CNRS, IMB, UMR, 5251, 33400 Talence, France

JD, 0000-0002-8335-9240; QG, 0000-0001-5978-9358; PM , 0000-0002-4776-0061

The infectiousness of infected individuals is known to depend on the time since the individual was infected, called the age of infection. Here, we study the parameter identifiability of the Kermack–McKendrick model with age of infection which takes into account this dependency. By considering a single cohort of individuals, we show that the daily reproduction number can be obtained by solving a Volterra integral equation that depends on the flow of newly infected individuals. We test the consistency of the method by generating data from deterministic and stochastic numerical simulations. Finally, we apply our method to a dataset from SARS-CoV-1 with detailed information on a single cluster of patients. We stress the necessity of taking into account the initial data in the analysis to ensure the identifiability of the problem.

1. Introduction

The existence of an individual infection (or contagiousness) period of variable length and magnitude among infected individuals is a proven fact in all contagious diseases. The origin of this variability is multiple. It may be due to (i) a variation in the symptomatic state of the infected person due to variable immune defences since the beginning of his infection; (ii) a variation in the environmental conditions of transport and survival of the infectious agent in the atmosphere, in a more or less favourable socio-sanitary environment presenting different spreading characteristics; (iii) a variation in the state of defence of the final host; (iv) a variation in the virulence of the infectious agent, which can mutate or possibly change the intermediary host; (v) a modification of the site of virus replication in the infected host and, therefore, a variation of the pathogen's transmissibility. In this article, we revisit the classical Kermack–McKendrick epidemic model with age of infection which takes into account a variability in the contagiousness of the hosts depending on the time lapse since the host was infected (case (i) above) and possibly the environmental conditions (case (ii) above). We develop a method to identify the individual transmissibility as a function of the age of infection.

(a) Continuous time model

Recall that the age of infection a is the time since individuals become infected. The major difficulty in matching the data and the Kermack–McKendrick model with age of infection is to identify: (i) the initial distribution of infected individuals with respect to the age of infection; (ii) the daily reproduction number $R_0(a)$ which is the reproduction number at the age of infection a (i.e. the average number of secondary cases produced by a single infected individual at the age of infection a). We can decompose the daily reproduction number as follows:

$$R_0(a) = \underbrace{\tau_0}_{\text{(A)}} \times \underbrace{S_0}_{\text{(B)}} \times \underbrace{\beta(a)}_{\text{(C)}} \times \underbrace{e^{-\nu a}}_{\text{(D)}},$$

where (A) τ_0 is the transmission rate at time t_0 (we assume the transmission rate to be constant during the period where $R_0(a)$ is evaluated). (B) S_0 is the average number of susceptible individuals at time t_0 with which an infected person may come into contact (we assume the number of susceptible individuals to be constant during the period where $R_0(a)$ is evaluated). (C) $\beta(a)$ is the probability of being infectious (i.e. capable of transmitting the pathogen) for an infected individual with age of infection a days. (D) $e^{-\nu a}$ is the probability for an infected individual with age of infection a days to remain infected.

Then the basic reproduction number (i.e. the number of secondary cases produced by a single infected individual) is given by

$$R_0 = \int_0^\infty R_0(a)\,\mathrm{d}a.$$

Here, we partly solve the problem of finding the initial distribution of infected patients by assuming that we start the epidemic at time t_0 with a single cohort of I_0 new infected patients. That is, the epidemic starts with I_0 infected patients all with age of infection $a = 0$. The case of an epidemic starting from a single infected patient (usually called the patient 0) corresponds to the case $I_0 = 1$. This is a common assumption in epidemiology. Note that the time t_0 at which the first patient becomes infected is also unknown for most epidemics.

Assume that the epidemic starts at time t_0 with a cohort of I_0 newly infected patients (i.e. all with age of infection $a = 0$). Then $N(t)$ the flow of new infected at time t satisfies the model starting from a single cohort of infected

$$N(t) = \underbrace{R_0(t - t_0) \times I_0}_{\text{(I)}} + \underbrace{\int_0^{t-t_0} R_0(s) \times N(t - s)\,\mathrm{d}s}_{\text{(II)}},\ \forall t \geq t_0, \tag{1.1}$$

where (I) is the flow of infected individuals at time t produced directly by the I_0 infected individuals already present on day t_0; and (II) is the flow of newly infected individuals at time t produced by the new infected individuals since day t_0.

The terminology 'flow of new infected individuals' means that the integral

$$\int_{t_1}^{t_2} N(\sigma)\,\mathrm{d}\sigma,$$

is the number of new infected individuals during the period of time $[t_1, t_2]$.

The Kermack–McKendrick model with age of infection is well defined only for integrable initial distribution. Equation (1.1) extends the Kermack–McKendrick model whenever the epidemic starts with a single cohort of infected, which corresponds to a Dirac mass initial distribution. This model remains valid as long as the transmission rate $\tau(t)$ and the number of susceptible hosts $S(t)$ remain constant with $\tau(t_0) = \tau_0$ and $S(t_0) = S_0$. So this model is valid when the epidemic starts.

Assume that I_0 is fixed and the function $a \to R_0(a)$ is given. Then the map $t \to N(t)$ can be obtained by solving (1.1). The goal of the article is to consider the converse problem. That is, assume that I_0 is fixed and assume that $t \to N(t)$ is given from the data. Then the map $a \to R_0(a)$ can be obtained by solving the Volterra integral equation

$$R_0(a) = \frac{N(a+t_0)}{I_0} - \frac{1}{I_0}\int_0^a R_0(s)N(a-s+t_0)\,\mathrm{d}s,\ \forall a \geq 0. \tag{1.2}$$

Therefore, if the map $t \to N(t)$ is known, we can theoretically derive the average dynamics of infection at the level of a single patient.

In this paper, we consider the whole time evolution of (1.1) starting from time $t = t_0$. That is in opposition to what is generally used in the literature. Indeed, people usually neglect the early beginning of the epidemic to consider the long-term evolution and assume that

$$R_0(a) = 0, \forall a \geq a_+, \tag{1.3}$$

where $a_+ > 0$ the maximal age of infectiousness for an infected patient. This leads to

$$N(t) = \int_0^{a_+} R_0(s) \times N(t-s)\,\mathrm{d}s, \quad \text{for } t \geq t_0 + a_+, \tag{1.4}$$

from which $a \to R_0(a)$ cannot be identified when $t \to N(t)$ is given. Indeed, assume for example that $N(t) = N_0\,\mathrm{e}^{\lambda t}$ is a given function. We obtain from (1.4) and after simplifications a standard characteristic equation

$$1 = \int_0^{a_+} R_0(s) \times \mathrm{e}^{-\lambda s}\,\mathrm{d}s. \tag{1.5}$$

The real number $\lambda > 0$ being given, if we consider $a \to \chi(a)$ any non-negative and non-null continuous function satisfying

$$\chi(a) = 0, \quad \forall a \geq a_+.$$

Then

$$R_0(a) = \frac{\chi(a)}{\int_0^{a_+} \chi(s) \times \mathrm{e}^{-\lambda s}\,\mathrm{d}s}$$

satisfies (1.5). Therefore, neglecting the initial value (I) in the Volterra equation (1.1) leads to a non-identifiable problem (in general). This shows the crucial role of the initial value in identifying the function $a \to R_0(a)$.

(b) Day by day model

The model (1.1) with a single cohort of infected becomes a discrete Volterra equation

$$N(t) = \underbrace{R_0(t-t_0) \times I_0}_{\text{(I)}} + \underbrace{\sum_{d=1}^{t-t_0} R_0(d) \times N(t-d)}_{\text{(II)}}, \quad \forall t \geq t_0, \tag{1.6}$$

where (I) is the number of infected produced directly by the I_0 infected individuals already present on day t_0; and (II) is the number of newly infected individuals at time t produced by the new infected individuals since day t_0.

Next, by setting $a = t - t_0$, we obtain the day-by-day equation for the daily reproduction number

$$R_0(a) = \frac{N(t_0+a)}{I_0} - \frac{1}{I_0}\sum_{d=1}^{a} R_0(d) \times N(t_0+a-d), \quad \forall a \geq 0. \tag{1.7}$$

In the above formula and throughout the paper, we use the following convention for the sum:

$$\sum_{d=k}^{m} = 0, \quad \text{whenever } m < k.$$

In practice, we can assume that $R_0(0) = 0$ since infected individuals are not infectious immediately after being infected. Under this additional assumption, we obtain the system

$$\begin{aligned}
N(t_0) &= 0,\\
N(t_0+1) &= R_0(1) \times I_0,\\
N(t_0+2) &= R_0(2) \times I_0 + R_0(1) \times N(t_0+1),\\
N(t_0+3) &= R_0(3) \times I_0 + R_0(2) \times N(t_0+1) + R_0(1) \times N_0(t_0+2),\\
&\vdots
\end{aligned}$$

When reliable information is available on the first cluster(s), the best formula for calculating daily basic reproduction numbers is equation (1.2) (or its discrete time version (1.7)). Based on (1.4), some methods have been developed in the literature to cope with the lack of precise information.

For instance, in [1,2], the authors following [3] propose an optimization algorithm for estimating the daily basic reproduction numbers. Unlike in the present article, the focus in [1,2] is on the variability with respect to time, not age of infection. In [4], D. Bernoulli mentions in 1760 the changes in the contagiousness parameters and places as a crucial challenge for the prediction of the transition between endemic and epidemic peaks in a prophetic sentence: 'Le retour d'une épidémie longtemps suspendue fait un ravage plus terrible dans une seule année qu'une endémie uniforme ne pourrait faire pendant un nombre d'années considérable' (The return of a long-suspended epidemic wreaks more terrible havoc in a single year than a uniform endemic could do for a considerable number of years). In [5], the authors use a deconvolution algorithm for calculating the daily basic reproduction numbers. In each case, the problem of the initial conditions is evoked at best only through the hypothesis of a unique 'patient zero'. Despite the considerable means of current investigation, in particular, those of the WHO and the members of the government of the WHO, it is rare that this patient is identified (this was the case for H1N1 in Mexico). The patient zero, also called index or primary case, is the first patient identified in a given population during an epidemiological investigation. It points out the source of the spread of a disease in a given reservoir, but this search is in general very difficult as was the case for HIV in North America [6].

One of the main difficulties in estimating the $R_0(a)$ function is its non-identifiability in general. Recent studies [7–11] developed methods to identify the various parameters for the COVID-19 pandemic by using cumulative reported cases data and differential equations models. Differential equations can be written in the form studied here by assuming that $R_0(a)$ (or equivalently, $\beta(a)$) is independent of a. Suppose that we are restricted to a period when the data is growing exponentially fast. If we take a fixed function $\beta(a)$, then by adapting the method developed in [11], we could identify a transmission rate τ so that the output of the model stays very close to the data, *for any function* $\beta(a)$. The same could be achieved with a good phenomenological description of the data by using the method developed in [7,9,10] with a time-dependent transmission rate. This means that the reported cases data is not sufficient to determine accurately the function $R_0(a)$. Without a good description of the initial distribution, it is hopeless to identify $R_0(a)$ by using reported cases only. We also refer to [12–14] for more indentification results. We also refer to [15] for an alternative method to estimate the basic reproduction number for a model with age of infection.

In this article, we first extend the Kermack–McKendrick model to initial conditions that are a linear combination of Dirac masses. The Kermack–McKendrick model cannot be extended in the space of measures (due to a lack of time continuity for the solutions). However, the Volterra integral equation can be extended and still makes perfect sense whenever we use Dirac masses for the initial condition. In many real examples, the initial distribution must be a linear combination of Dirac masses since the data are discrete at the early stage of an epidemic (in a city, a country). Indeed, at the early beginning of an epidemic, the epidemic starts from a few cases imported from other places. Therefore, Dirac masses make perfect sense. In practice, the early stage of an outbreak is often undocumented and generally difficult to determine. But our study applies to data from a finite collection of clusters that is easier to determine using contact tracing. By clusters, we mean the descendants produced by direct or indirect contact from a finite number of infected individuals (locally concentrated in space).

Consequently, for the single cohort model, we can reverse the problem, and by assuming that the daily number of new infected is known, we can compute the daily reproduction number by solving a Volterra integral equation. The daily basic reproduction number informs us about the dynamics of infection at the level of a single patient. Therefore, knowing $R_0(a)$ should help the medical doctors decide about quarantine measures. Reported case data for clusters are particularly valuable for reconstructing the dynamics of infection at the level of a single individual.

In this paper, we also provide an individual-based model (IBM) (see electronic supplementary material). This IBM converges to the deterministic model whenever the initial number of infected increases. We use this IBM to generate sample data to test our method and compute the daily basic reproduction number. This will allow us to test the effects of the day-to-day discretization (on the data) and the impact of stochastic perturbations on the daily reproduction numbers. We conclude the paper by applying our approach to a cluster of SARS-CoV-1 in Singapore.

The plan of the paper is the following. In §2, we recall the Kermack–McKendrick model with age of infection. We explain how to derive the Volterra formulation of the model, and we compare it with the Kermack–McKendrick SI model with age of infection (ODE model). In §3, we explain how to connect the model with the data. In §4, we extend the Kermack–McKendrick model with age of infection in the case where the epidemic starts from a single or multiple cohorts of infected individuals. In §5, we derive an equation to compute the daily reproduction number from the data. In §6, we consider a day by day discretized Kermack–McKendrick model with age of infection. In §7, we run some numerical simulations, we compare the deterministic model with a stochastic individual-based simulation presented in electronic supplementary material. In §8, we compare the model with some data from SARS-CoV-1, and we discuss the data from SARS-CoV-2.

2. Kermack–McKendrick model with age of infection

(a) Partial differential equation formulation of the model

The age of infection a is the time since individuals become infected. Let $a \to i(t,a)$ be the distribution of population of *infected individuals* at time t (with respect to a the age of infection). The term distribution of population means that the integral

$$\int_{a_1}^{a_2} i(t,a)\,\mathrm{d}a$$

is the number of infected at time t with infection age between a_1 and a_2. Therefore, the total number of infected individuals at time t is

$$I(t) = \int_0^{+\infty} i(t,a)\,\mathrm{d}a.$$

Let $\beta(a) \in [0,1]$ be the probability to be contagious or infectious (i.e. capable of transmitting the pathogen) at the age of infection a. The quantity $\beta(a)$ can be interpreted as the fraction of infected individuals with age of infection a that are infectious. Then the total number of *contagious individuals* (or also called *infectious individuals*; i.e. the individuals capable of transmitting the pathogen) at time t is

$$C(t) = \int_0^{+\infty} \beta(a) i(t,a)\,\mathrm{d}a.$$

The model of Kermack–McKendrick [16] with age of infection is the following, for each $t \geq t_0$

$$\left.\begin{aligned} S'(t) &= -\tau(t)\,S(t)\int_0^{+\infty} \beta(a)\,i(t,a)\,\mathrm{d}a \\ \partial_t i + \partial_a i &= -\nu\, i(t,a), \quad \text{for } a \geq 0 \\ \text{and} \qquad i(t,0) &= \tau(t)\,S(t)\int_0^{+\infty} \beta(a)\,i(t,a)\,\mathrm{d}a, \end{aligned}\right\} \tag{2.1}$$

this system is supplemented by initial data

$$S(t_0) = S_0 \geq 0 \quad \text{and} \quad i(t_0,a) = i_0(a) \in L^1_+(0,\infty), \tag{2.2}$$

where $L^1_+(0,\infty)$ is the positive cone of non-negative integral function.

In the model, $S(t)$ is the number of susceptible individuals at time t, $t \to \tau(t)$ is the transmission rate at time t, and $\nu \geq 0$ is the rate at which individuals die or recover. The time changes of the transmission rate $\tau(t)$ is the combination of three factors. First, the coefficient of virulence, linked to the infectious agent. The coefficient of virulence may change over time due to mutations of the pathogen. Second, the coefficient of susceptibility, linked to the host. These two first factors are all summarized into the probability of transmission. Third, the number of contacts per unit of time between individuals (this number is directly connected to the mitigation measures). Here, the parameter ν is assumed to be independent of the age of infection a. This is a simplifying assumption to improve the readability of the paper. The parameter ν combines both the specific fatality rate and the recovery rate.

The above equation can be understood first as follows:

$$I'(t) = \underbrace{\tau(t)\,S(t)\int_0^{+\infty} \beta(a)\,i(t,a)da}_{\text{(I)}} - \underbrace{\int_0^{+\infty} \nu\, i(t,a)\,\mathrm{d}a}_{\text{(II)}},$$

where (I) is the flow of new infected, and (II) is the flow of individuals who die or recover.

We make the following assumption.

Assumption 2.1. We assume that

(i) The transmission rate $t \to \tau(t)$ is a bounded continuous map from $[t_0, +\infty)$ in $[0, +\infty)$;
(ii) The probability to be infectious at the age of infection $a \to \beta(a) \in L^\infty_+(0, +\infty)$ is a non-negative and measurable function of a which is bounded by 1.

(b) Volterra integral equation formulation of the model

In the model (2.1), the quantity

$$N(t) := \tau(t)\, S(t) \int_0^{+\infty} \beta(a)\, i(t,a)\, \mathrm{d}a, \tag{2.3}$$

is the flow of new infected individuals at time t.

By using the S-equation in system (2.1), we obtain

$$S(t) = S_0 - \int_{t_0}^{t} N(\sigma)\, \mathrm{d}\sigma, \quad \forall t \geq t_0. \tag{2.4}$$

By integrating the second equation of system (2.1) along the characteristics, we obtain

$$i(t,a) = \begin{cases} \mathrm{e}^{-\nu\,(t-t_0)}\, i_0(a-(t-t_0)), & \text{if } a \geq t - t_0, \\ \mathrm{e}^{-\nu\, a} N(t-a), & \text{if } t - t_0 \geq a. \end{cases} \tag{2.5}$$

By using (2.5), we deduce that $t \to N(t)$ satisfies the following Volterra integral equation

$$N(t) = \underbrace{\tau(t)\, S(t) \int_{t-t_0}^{+\infty} \beta(a)\, \mathrm{e}^{-\nu\,(t-t_0)} i_0(a-(t-t_0))\mathrm{d}a}_{\text{(I)}} + \underbrace{\tau(t)\, S(t) \int_0^{t-t_0} \beta(a)\, \mathrm{e}^{-\nu\, a} N(t-a)\, \mathrm{d}a}_{\text{(II)}}, \tag{2.6}$$

where (I) is the flow of newly infected individuals at time t produced by the infected individuals already present on day t_0; (II) is the flow of newly infected individuals at time t produced by the newly infected individuals since day t_0.

By using equations (2.4) and (2.6), we can summarize the epidemic model (2.1), by saying that $t \to N(t)$ is the unique continuous map satisfying

$$N(t) = \tau(t)\, S(t) \left[\Lambda(t) + \int_0^{t-t_0} \beta(a)\, \mathrm{e}^{-\nu\, a} N(t-a)\, \mathrm{d}a \right], \quad \forall t \geq t_0, \tag{2.7}$$

where

$$S(t) = S_0 - \int_{t_0}^{t} N(\sigma)\, \mathrm{d}\sigma, \forall t \geq t_0 \tag{2.8}$$

and

$$\Lambda(t) := \mathrm{e}^{-\nu(t-t_0)} \int_{t-t_0}^{+\infty} \beta(a)\, i_0(a-(t-t_0))\mathrm{d}a, \quad \forall t \geq t_0. \tag{2.9}$$

The function $\Lambda(t)$ is the number of infectious individuals (capable of transmitting the pathogen) at time t among the infected individuals already present at time t_0.

The function $t \to \Lambda(t)$ plays a fundamental role in solving the Volterra equation. Indeed, the quantity

$$\int_{t_1}^{t_2} \tau(\sigma)\, S(\sigma) \Lambda(\sigma)\, \mathrm{d}\sigma$$

is the number of infected produced between the instants t_1 and t_2 by the infected already present at time t_0. So, for example, if no new infected are produced by the infected already present at time t_0, that is if $\Lambda(t) = 0$, $\forall t \geq t_0$, then there will be no new infected at all after the time t_0, that is $N(t) = 0$, $\forall t \geq t_0$. The function $t \to \Lambda(t)$ can be regarded as the *initial condition (or initial distribution)* for the Volterra integral equation (2.7).

Remark 2.2. In the case of the standard *SI* model, which is

$$\begin{cases} S'(t) = -\tau(t)S(t)I(t), \\ I'(t) = \tau(t)S(t)I(t) - \nu I(t), \end{cases} \qquad \text{for } t \geq t_0.$$

By applying the variation of constant formula to I-equation, we obtain

$$I(t) = \mathrm{e}^{-\nu(t-t_0)}I_0 + \int_{t_0}^{t} \mathrm{e}^{-\nu(t-s)}N(s)\,\mathrm{d}s.$$

Therefore, by replacing $I(t)$ by the above formula in the equation $N(t) = \tau(t)S(t)I(t)$, we obtain a Volterra integral equation for N

$$N(t) = \underbrace{\tau(t)\,S(t)\,\mathrm{e}^{-\nu\,(t-t_0)}I_0}_{\text{(I)}} + \underbrace{\tau(t)\,S(t)\int_0^{t-t_0} \mathrm{e}^{-\nu\,a}N(t-a)\,\mathrm{d}a}_{\text{(II)}}. \tag{2.10}$$

We conclude that the above Volterra integral equation corresponds to (2.7) in the special case where $\beta(a) = 1$, for almost every $a \geq 0$. From (2.10), it becomes clear that $\tau(t)\,S(t)\,\Lambda(t) = \tau(t)\,S(t)\,\mathrm{e}^{-\nu\,(t-t_0)}I_0$ is the contribution to the flow of newly infected individuals at time t produced by the I_0 infected individuals already present at time t_0.

3. Connecting the data and the model

The data are represented by the function $t \to \mathrm{CR}(t)$ which is the cumulative number of reported cases at time t. We propose as a model that the flow of reported cases is a fraction $0 \leq f \leq 1$ of the flow of recovering individuals, that is

$$\mathrm{CR}'(t) = f\,\nu \int_0^{+\infty} i(t,a)\,\mathrm{d}a. \tag{3.1}$$

By using (2.5), we can compute the number of infected at time t. That is

$$\int_0^{+\infty} i(t,a)\,\mathrm{d}a = \mathrm{e}^{-\nu\,(t-t_0)}I_0 + \int_0^{t-t_0} \mathrm{e}^{-\nu\,a}N(t-a)\,\mathrm{d}a, \tag{3.2}$$

where

$$I_0 = \int_0^{+\infty} i_0(a)\,\mathrm{d}a$$

is the total number of infected at time t_0.

By using equations (3.1) and (3.2), we obtain

$$\mathrm{CR}'(t) = f\nu\left[\mathrm{e}^{-\nu\,(t-t_0)}I_0 + \int_0^{t-t_0} \mathrm{e}^{-\nu\,a}N(t-a)\,\mathrm{d}a\right],$$

or equivalently (by using the change of variable $\sigma = t - a$)

$$\mathrm{CR}'(t) = f\nu\left[\mathrm{e}^{-\nu\,(t-t_0)}I_0 + \int_{t_0}^{t} \mathrm{e}^{-\nu\,(t-\sigma)}N(\sigma)\,\mathrm{d}\sigma\right].$$

By choosing $t = t_0$, we obtain

$$I_0 = \frac{\mathrm{CR}'(t_0)}{f\nu}$$

and

$$\int_{t_0}^{t} \mathrm{e}^{\nu\sigma}N(\sigma)\,\mathrm{d}\sigma = \frac{\mathrm{e}^{\nu\,t}\mathrm{CR}'(t)}{f\nu} - \mathrm{e}^{\nu\,t_0}I_0,$$

and by differentiating both sides of the above equation, we obtain

$$\mathrm{e}^{\nu t}N(t) = \frac{\nu\,\mathrm{e}^{\nu\,t}\mathrm{CR}'(t) + \mathrm{e}^{\nu\,t}\mathrm{CR}''(t)}{f\nu}.$$

Therefore, we obtain the following connection between the data and the model.

Connection between the data and the model

Let $t \to \mathrm{CR}(t)$ be the cumulative number of reported cases. Then the initial number of infected is given by

$$I_0 = \frac{\mathrm{CR}'(t_0)}{f\nu}, \tag{3.3}$$

and the flow of new infected individuals $N(t)$ at time t is given by

$$N(t) = \frac{\nu \mathrm{CR}'(t) + \mathrm{CR}''(t)}{f\nu}, \quad \forall t \geq t_0. \tag{3.4}$$

Remark 3.1. In practice, it is possible but not easy to have a reliable evaluation of $t \to \mathrm{CR}'(t)$ and especially $t \to \mathrm{CR}''(t)$. This problem was considered by using some averaging (or phenomenological models) procedure of the reported sanitary data [7,9,10,17].

4. Kermack–McKendrick model starting from a single and multiple cohorts of infected patients

The major difficulty to compare the model (2.4) with the data is to identify the functions $a \to i_0(a)$ and $a \to \beta(a)$. To simplify the discussion, let us consider the model at the early stage of the epidemic. When the epidemic just starts we can assume that the transmission rate $t \to \tau(t)$ remains constant, and the number of susceptible individuals $t \to S(t)$ is constant and equal to S_0. Under such a simplifying assumption the Volterra equation (2.4) becomes

$$N(t) = \tau\, S_0 \left[\Lambda(t) + \int_0^{t-t_0} \beta(a)\, \mathrm{e}^{-\nu a} N(t-a)\, \mathrm{d}a \right], \quad \forall t \geq t_0. \tag{4.1}$$

(a) A single cohort initial distribution for the PDE model

In order to understand the mathematical concept of Dirac mass centered at 0, we first consider an approximation by an exponential law

$$i_0(a) = I_0\, \kappa\, \mathrm{e}^{-\kappa a}, \tag{4.2}$$

with mean and standard deviation equal to $1/\kappa$. Then a Dirac mass centered at age 0 can be understood as the limit of such a distribution when κ goes to $+\infty$. The limit needs some explanations. Recall that

$$\int_{a_1}^{a_2} i_0(a)\, \mathrm{d}a = I_0 \left[\mathrm{e}^{-\kappa a_1} - \mathrm{e}^{-\kappa a_2} \right]$$

is the initial number of infected individuals with infection age a in between a_1 and a_2 at time $t = 0$. We deduce that

$$\lim_{\kappa \to \infty} \int_{a_1}^{a_2} i_0(a)\, \mathrm{d}a = \begin{cases} 0, & \text{if } a_2 > a_1 > 0, \\ I_0, & \text{if } a_2 > a_1 = 0. \end{cases}$$

That is to say that, when κ tends to $+\infty$, the initial distribution of population $i_0(a)$ is approaching the case where all the infected individuals at time t_0 have the same age of infection $a = 0$.

For short, we write

$$i_0(a) = I_0\, \delta_0(a),$$

where $\delta_0(a)$ is called the Dirac mass centered at age 0.

(b) A single cohort initial distribution for the Volterra integral equation

Recall that

$$\Lambda(t) = \mathrm{e}^{-\nu\,(t-t_0)} \int_0^{+\infty} \beta(a + (t-t_0))\, i_0(a)\, \mathrm{d}a,$$

so when $i_0(a)$ is replaced by (4.2) (with an explicit dependency on κ) we obtain

$$\Lambda_\kappa(t) := I_0\, \mathrm{e}^{-\nu\,(t-t_0)} \int_0^{+\infty} \beta(a + (t-t_0))\, \kappa\, \mathrm{e}^{-\kappa a}\, \mathrm{d}a.$$

From now on, every function that depends on κ will be indexed by κ.

In order to derive the Kermack–McKendrick model with Dirac mass initial distribution as limit, we first need the following result. The proof of the following can be found in the electronic supplementary material.

Lemma 4.1. *Let assumption 2.1 be satisfied, and assume in addition that $a \to \beta(a)$ is continuous. Then we have*

$$\lim_{\kappa\to\infty} \Lambda_\kappa(t) = I_0\, \mathrm{e}^{-\nu\,(t-t_0)} \beta(t-t_0),$$

where the limit is uniform in $t \geq t_0$. That is

$$\lim_{\kappa\to+\infty} \sup_{t\geq t_0} |\Lambda_\kappa(t) - I_0 e^{-\nu\,(t-t_0)} \beta(t-t_0)| = 0.$$

The initial condition of the Volterra integral equation (2.7) becomes at the limit $\Lambda(t) = I_0\, \mathrm{e}^{-\nu\,(t-t_0)} \beta(t-t_0)$. One may observe that the above limit can be obtained for many types of approximation of the Dirac mass centered at 0 (probability distribution on $(0, +\infty)$). So formula (4.2) can be replaced by another formula.

(c) A single cohort Volterra integral equation model

Define

$$\Gamma(a) = \mathrm{e}^{-\nu\, a} \beta(a), \quad \forall a \geq 0. \tag{4.3}$$

Then by using (2.6), the Kermack–McKendrick model can be reformulated for $t \geq t_0$, as the following system

$$N_\kappa(t) = \tau(t) S_\kappa(t) \left[\Lambda_\kappa(t) + \int_0^{t-t_0} \Gamma(a)\, N_\kappa(t-a)\, \mathrm{d}a \right],$$

where $\Lambda_\kappa(t)$ is defined above, and

$$S_\kappa(t) = S_0 - \int_{t_0}^{t} N_\kappa(\sigma)\, \mathrm{d}\sigma.$$

By taking first a formal limit when $\kappa \to +\infty$, we obtain the model starting from a single cohort of infected.

Kermack–McKendrick model starting from a single cohort of infected

Assume that the initial distribution of infected only contains a single cohort composed of I_0 individuals all with age of infection $a = 0$ at time t_0. Then the flow of new infected $t \to N(t)$ is the unique continuous solution of the Volterra integral equation

$$N(t) = \tau(t) S(t) \left[I_0 \times \Gamma(t-t_0) + \int_0^{t-t_0} \Gamma(a)\, N(t-a)\, \mathrm{d}a \right], \quad \forall t \geq t_0, \tag{4.4}$$

where $S(t)$ is obtained from (2.4).

The following theorem says that the model with a single cohort of infected extends the earlier model of Kermack–McKendrick with initial distribution in L^1. This theorem is a consequence of

lemma 4.1 and of the continuity of the semiflow generated by the Volterra integral equation. We refer to Ducrot and Magal [18] for more results on this topic.

Theorem 4.2. *Let assumption 2.1 be satisfied, and assume in addition that $a \to \beta(a)$ is continuous. Then*

$$\lim_{\kappa \to \infty} N_\kappa(t) = N(t),$$

where the limit is uniform in t on every closed and bounded interval of $[t_0, +\infty)$, and the map $t \to N(t)$ is the unique continuous solution of the Volterra integral equation (4.4).

Remark 4.3. When the initial distribution is a Dirac mass centered at $a = 0$, the total number of infected individuals at time t is

$$C(t) = \mathrm{e}^{-\nu(t-t_0)} I_0 + \int_0^{t-t_0} \mathrm{e}^{-\nu a} N(t-a)\,\mathrm{d}a, \quad \forall t \geq t_0,$$

and the number of infectious individuals at time t is

$$I(t) = \beta(t-t_0)\mathrm{e}^{-\nu(t-t_0)} I_0 + \int_0^{t-t_0} \beta(a)\,\mathrm{e}^{-\nu a} N(t-a)\,\mathrm{d}a, \quad \forall t \geq t_0.$$

In the case of multiple cohorts, the initial distribution becomes $i_0(a) = I_0^1\,\delta_{a_1}(a) + \ldots + I_0^n\,\delta_{a_n}(a)$, where $a_1 < a_2 < \ldots < a_n$ are the ages of infection for each cohort at time t_0, and I_0^j is the number of infected in the jth-cohort at time t_0. By analogy to the case of a single cohort, we can approach the initial condition as follows:

$$\Lambda_\kappa(t) := \sum_{j=1}^{n} I_0^j\,\mathrm{e}^{-\nu(t-t_0)} \int_0^{+\infty} \beta(a + (t-t_0))\,\kappa\,\mathrm{e}^{-\kappa(a-a_j)} \mathbb{1}(a-a_j)\mathrm{d}a;$$

and by using (4.3), we obtain

$$\Lambda_\kappa(t) \to \sum_{j=1}^{n} \mathrm{e}^{-\nu(t-t_0)} \beta((t-t_0)+a_j) I_0^j = \sum_{j=1}^{n} \Gamma(t-t_0+a_j) \frac{I_0^j}{\mathrm{e}^{-\nu a_j}},$$

when $\kappa \to +\infty$.

Kermack–McKendrick model starting from multiple cohorts of infected

Assume that the initial distribution of infected consists of $n \geq 1$ cohorts of infected with age of infection $a_1 < a_2 < \ldots < a_n$ at time t_0. That is

$$i_0(a) = I_0^1\,\delta_{a_1}(a) + \ldots + I_0^n\,\delta_{a_n}(a),$$

where I_0^j is the number of infected in the jth-cohort at time t_0.

Then the flow of infected $t \to N(t)$ satisfies the following Volterra integral equation:

$$N(t) = \tau(t) S(t) \left[\sum_{j=1}^{n} \Gamma(t-t_0+a_j) \frac{I_0^j}{\mathrm{e}^{-\nu a_j}} + \int_0^{t-t_0} \Gamma(a)\,N(t-a)\,\mathrm{d}a \right],$$

where $S(t)$ is obtained from (2.4).

(d) Basic reproduction number

In this section, we assume that the transmission $t \to \tau(t)$ is constant equal to τ, and $t \to S(t)$ is constant equal to S_0.

Define the **daily reproduction numbers**

$$R_0(a) = \tau \times S_0 \times \Gamma(a) = \tau \times S_0 \times \beta(a) \times \mathrm{e}^{-\nu a}, \quad \forall a \geq 0. \tag{4.5}$$

Basic reproduction number

The total number of the first generation of newly infected produced by a single infected patient with age of infection $a = 0$ at time $t = t_0$ is called the **basic reproduction number**. That is

$$R_0 = \int_0^\infty R_0(a)\,\mathrm{d}a.$$

The flow of the first generation of newly infected produced by a single infected patient who has been infected for a days is called the **daily reproduction numbers**. When the time unit is 1 day, the function $R_0(a)$ is also the average daily number of cases produced by a single patient at the age of infection a.

Remark 4.4. The total number of cases produced by the nth generation of infected resulting from a single infected patient is

$$\int_0^\infty (R_0^{*(n)})(t)\,\mathrm{d}t = (R_0)^n.$$

5. Computing the age dependent reproduction number $\Gamma(a)$ from the data

By using (4.4), one may realize that instead of computing $N(t)$ as a function of $\Gamma(a)$, we can reverse the computations and obtain $\Gamma(a)$ as a function of $N(t)$ (regarded as the data).

Computing $\Gamma(a)$ from the data

Assume in addition that the parameters t_0, $S_0 > 0$, I_0, $\nu > 0$, and the function $t \to \tau(t)$ are known. Then the function $t \to \Gamma(t)$ can be obtained from the flow of newly infected $t \to N(t)$, as the unique solution of the Volterra integral equation

$$\Gamma(t - t_0) = \frac{1}{I_0}\left(\frac{N(t)}{\tau(t)S(t)} - \int_0^{t-t_0} \Gamma(a)\,N(t-a)\,\mathrm{d}a\right), \quad \forall t \geq t_0, \tag{5.1}$$

where $S(t)$ is obtained by using (2.4).

Remark 5.1. Assume that patients cannot transmit the pathogen when the age of infection is above $a^+ > 0$. That is

$$\Gamma(a) = 0, \quad \forall a \geq a^+.$$

Then equation (5.1) becomes for all $t \geq t_0 + a^+$,

$$\frac{N(t)}{\tau(t)S(t)} = \int_0^{a^+} \Gamma(a)\,N(t-a)\,\mathrm{d}a \Leftrightarrow N(t) = \tau(t)S(t)\int_0^{a^+} \Gamma(a)\,N(t-a)\,\mathrm{d}a.$$

6. Day by day Kermack–McKendrick model with age of infection

The variation of the number of susceptible individuals $S(t)$ is given each day $t = t_0, t_0 + 1, \ldots$, by

$$S(t) = S_0 - \sum_{d=t_0}^{t-1} N(d), \tag{6.1}$$

where S_0 is the number of susceptible on day 0, $S(t)$ is the number of susceptible on day t and $N(d)$ is the daily number of new infected individuals on day d. By analogy with equation (2.6), the daily number of newly infected individuals satisfies the following discrete time Volterra integral

equation for all $\forall t = t_0, t_0+1, t_0+2, \ldots,$

$$N(t) = \tau(t)S(t)\sum_{d=t-t_0}^{+\infty} \Gamma(d)\frac{I_0(d-(t-t_0))}{e^{-\nu\,(d-(t-t_0))}} + \tau(t)S(t)\sum_{d=1}^{t-t_0}\Gamma(d)\times N(t-d), \tag{6.2}$$

where

$$\Gamma(d) := \beta(d)\,\mathrm{e}^{-\nu\,d}, \quad \forall d = 0, 1, 2, \ldots,$$

and $\tau(t)$ is the transmission rate, $\beta(d)$ is the probability of being infectious (i.e. capable of transmitting the pathogen) after d days of infection and $e^{-\nu\,d}$ is the probability of staying infected after d days of infection (i.e. the probability of neither recovering nor dying after d days of infection). The quantity $I_0(d)$ is the number of infected on day 0 which have been infected d days ago.

The model (4.4) with a single cohort of infected becomes

$$N(t) = \tau(t)S(t)\left[\Gamma(t-t_0)I_0(0) + \sum_{d=1}^{t-t_0}\Gamma(d)\times N(t-d)\right]. \tag{6.3}$$

Day by day single cohort model and daily basic reproduction number

Assume that $t \to \tau(t)$ equals τ_0, and $t \to S(t)$ is constant equal to S_0. Assume that the epidemic starts at time t_0 with a cohort of I_0 newly infected patients (i.e. with age of infection $a = 0$). The model with a single cohort of infected becomes a discrete Volterra equation

$$N(t) = \left[R_0(t-t_0)\times I_0 + \sum_{d=1}^{t-t_0} R_0(d)\times N(t-d)\right], \quad \forall t \geq t_0. \tag{6.4}$$

We obtain the day-by-day equation for the daily reproduction number

$$R_0(a) = \frac{N(t_0+a)}{I_0} - \frac{1}{I_0}\sum_{d=1}^{a} R_0(d)\times N(t_0+a-d), \quad \forall a \geq 0. \tag{6.5}$$

7. Numerical simulations

In the simulations, the unit of time is 1 day, and we fix

$$S_0 = 10^7 = 10\,000\,000, \quad \frac{1}{\nu} = 9\,\text{days, and } R_0 = 1.1.$$

For each function $\beta(a)$ described below, the parameter τ is obtained numerically by using the following formula

$$\tau = \frac{R_0}{S_0\int_0^\infty \beta(a)\,\mathrm{e}^{-\nu a}\,\mathrm{d}a},$$

where the integral is computed by using the Simpson integration method.

In the following, we use the numerical scheme described in electronic supplementary material to run the simulation of the Volterra integral equations (4.4).

(a) Stochastic simulations: individual-based model

In order to estimate the uncertainty expected in real datasets, we use stochastic simulations that reproduce the first stages of the epidemic in finite populations. We consider a population composed of a finite number $N = S_0 + I_0$ of individuals. We start the simulation at time $t = 0$ with $S_0 \in \mathbb{N}$ susceptible individuals and $I_0 \in \mathbb{N}$ infected individuals all with age of infection $a = 0$. For each infected individual, we also compute the time spent in the I-compartment which follows an exponential law with parameters $1/\nu$. The principles of the simulations are as follows:

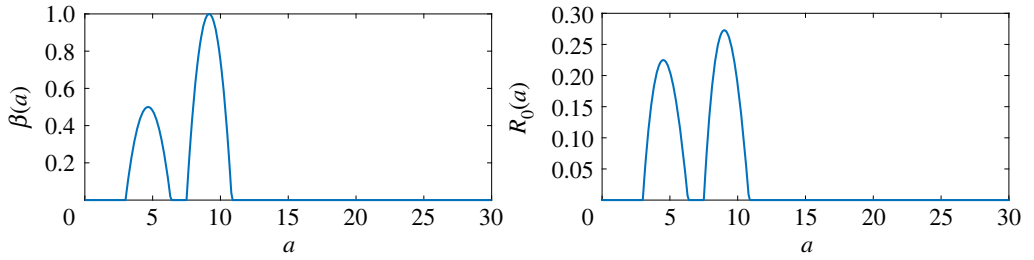

Figure 1. On the left-hand side, we plot the function $a \to \beta(a)$. On the right-hand side, we plot the function $a \to R_0(a) = \tau_0 \times S_0 \times \beta(a) \times e^{-\nu a}$.

(i) Individuals meet at random at rate $\tau > 0$. In other words, each pair of individuals in the population has a contact which occurs at a time following an exponential law of average $1/\tau$.

(ii) When a contact occurs between an infected individual of age a and a susceptible individual, the contact results in a newly infected individual of age 0 with probability $\beta(a)$. When the infection occurs, the newly infected individual is assigned a duration of infection which follows an exponential law of rate ν. Therefore, individuals stay infected on average for a duration of $1/\nu$.

(iii) The age of all individuals is updated at fixed intervals of time of size Δt. Simultaneously, the lifespan of each infected individual is decreased by Δt and individuals whose lifespan has become negative are removed from the system.

The MATLAB code of the IBM is available online at: https://github.com/romainvieme/2022-kermack-mckendrick-single-cohort.

(b) Numerical evidence of the convergence of the IBM to the deterministic model

In this section, we illustrate the convergence of the IBM to the deterministic model whenever I_0 increases.

It is common to see biphasic flu clinically: after incubation of 1 day, there is a high fever, then a drop in temperature before rising again, hence the term 'V' fever [19]. Such a biphasic contagiousness is also observed in COVID-19. The viral load in throat swab and sputum has been measured for COVID-19 patients, which leads to biphasic contagiousness [5,20]. To cover these types of infectious diseases, we introduce the following form for the probability to be infectious

$$\beta(a) = 0.5 \times 4q\{(a - a_0)(1 - q(a - a_0))\}^+ + 4q\{(a - pa_0)(1 - q(a - pa_0))\}^+, \tag{7.1}$$

with $a_0 = 3$ days, $p = 2.5$, and $q = 0.3$ (see figure 1).

First generation of secondary cases produced by a single infected: In figure 2, we use the IBM to investigate some properties of the clusters obtained from the stochastic simulations. We compare such a stochastic sample with the original $a \to R_0(a)$.

Figures 3–6 clearly show the influence of the distribution of daily reproduction numbers throughout the period of contagiousness, with distribution assumed to be identical for all infected individuals. When it is biphasic, our method makes it possible to estimate it with good precision using the IBM stochastic model (figure 2). As the IBM model converges towards the deterministic model when we increase the size of the simulated sample (figure 3), we can anticipate that the biphasic estimate remains precise for the deterministic version of the model, which is observed in figure 3. The large fluctuations observed in the IBM simulations of the daily reproduction numbers (figure 3) are indeed considerably attenuated if we consider the average curves corresponding to different samples of 500 IBM runs.

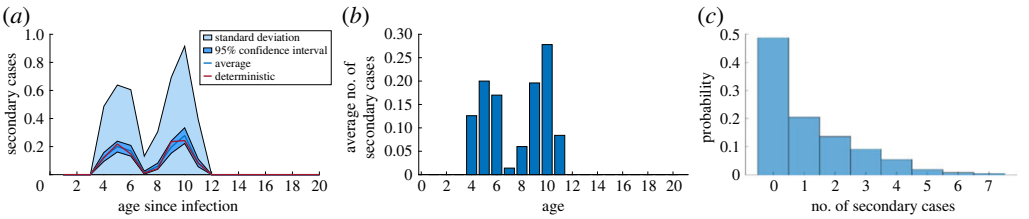

Figure 2. In these figures, we present sets of 500 samples of secondary cases produced by a single infected individual in a population of $S = 10^7$ susceptible hosts. Theses samples are produced by using the IBM. (*a*) Statistical summary: the blue curve represents the average number of cases at age of infection a; the dark blue area is the 95% confidence interval of this average obtained by fitting a Gaussian distribution to the data; the light blue area corresponds to the standard deviation; the orange curve is the deterministic daily basic reproductive number at age a. (*b*) Bar graph of the average number of secondary cases as a function of the age since infection. (*c*) Histogram of the total number of secondary cases produced during the whole infection. This estimates the probability of a single infected generating n secondary cases (with n in the abscissa).

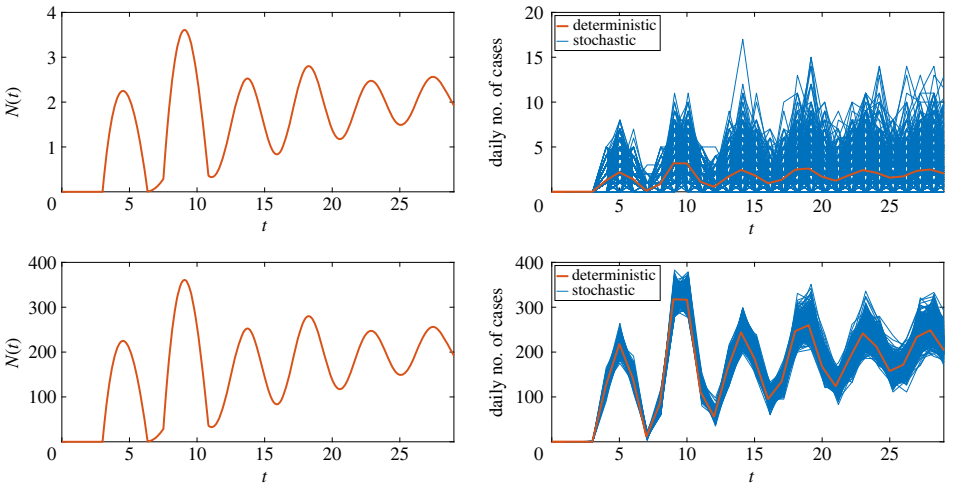

Figure 3. On the left-hand side, we plot the function $t \longrightarrow N(t)$ solution of (4.4) with (2.4). On the right-hand side, we plot the function $t \longrightarrow \int_{t-1}^{t} N(s)\,\mathrm{d}s$ (for $t = 1, 2, \ldots$) which corresponds to the daily number of cases obtained by solving (4.4) with (2.4), and we compare it with the daily number of cases obtained from 500 runs of the IBM. The top two figures correspond to $I_0 = 10$, and the bottom two figures to $I_0 = 1000$.

In figure 4, we focus on the reconstruction of the daily reproduction number from deterministic simulations. In figure 4, we observe the effect of the day-by-day discretization (which corresponds to the daily reported data). In figures 5 and 6, we focus on the reconstruction of the daily reproduction number from stochastic simulations. In figures 5 and 6, we observe the stochastic effect of the IBM.

It can be noted in figure 4 that the variations in the daily reproduction numbers of an individual are identical for a set of I_0 equal to $6, 10, 14$ and for a set equal to $600, 1000, 1400$. The reason for this similarity is related to the normalization of the simulated daily reproduction numbers by the size of the set of initial infected individuals, in order to reduce them to an individual. In figure 4 there is an important negativity in late daily reproduction numbers, when the duration of the period of contagiousness is high and the initial number of infected is small. This phenomenon is very attenuated in the stochastic model, if we take the average of many simulations of the IBM model, even in the case where the initial number of infected is small (figures 5 and 6).

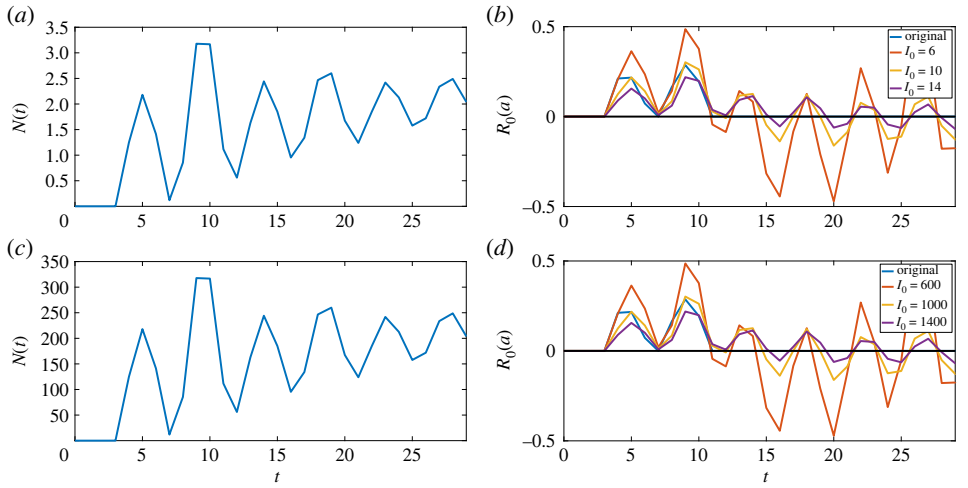

Figure 4. On the left-hand side, we plot the daily number of $t \to \int_{t-1}^{t} N(s)\,\mathrm{d}s$ (for $t = 0, 1, 2, \ldots$) by using the continuous model (1.1) for $I_0 = 10$ (*a*,*b*) and $I_0 = 1\,000$ (*c*,*d*). On the right-hand side, we apply formula (6.5) to the flow of new infected obtained from the deterministic model. In the top two figures, we vary $I_0 = 6, 10, 14$. In the bottom two figures, we vary $I_0 = 600, 1000, 1400$. In both cases, the yellow curve gives the best visual fit, and the $R_0(a)$ becomes negative whenever I_0 becomes too small.

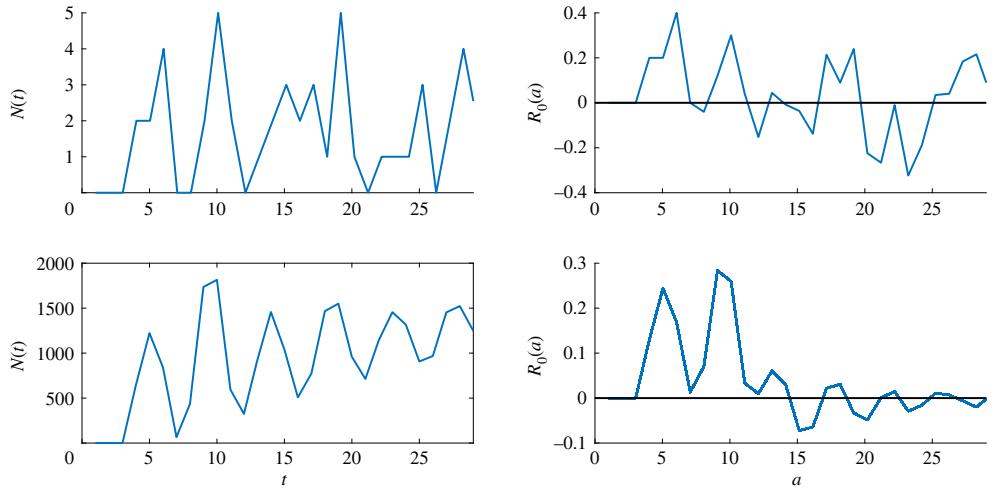

Figure 5. On the left-hand side, we plot the daily number of cases $t \to N(t)$ (for $t = 0, 1, 2, \ldots$) obtained on the top from a single run of the IBM, and the bottom by summing the daily number of cases for 500 IMB runs. On the right-hand side, we apply formula (6.5) (with $I_0 = 10$) to the daily number of cases obtained from the IBM. The top two figures correspond to $I_0 = 10$, and the bottom two figures to $I_0 = 500 \times 10$.

8. Application to SARS-CoV-1

In practice, the Kermack–McKendrick model starting from a Dirac mass means that the epidemic starts from a single patient at time t_0 (whenever $I_0 = 1$) or from a group of I_0 infected patients all with the same age of infection $a = 0$ at time t_0. This assumption corresponds to the standard conception of a cluster in epidemiology. An example of such a cluster is obtained [21] for the SARS-CoV-1 epidemic in Singapore in 2003. This cluster is represented by a network of contacts between individuals in figure 7.

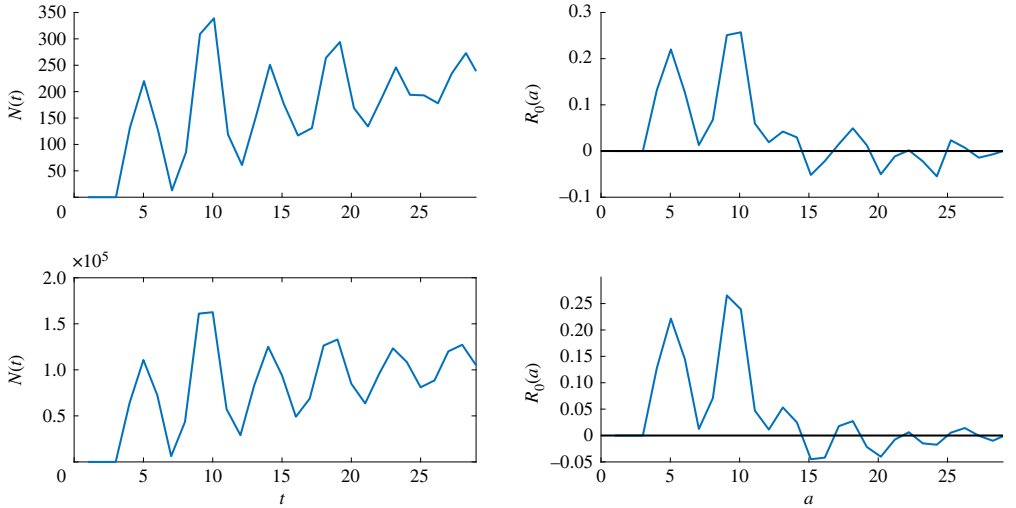

Figure 6. On the left-hand side, we plot the daily number of cases $t \to N(t)$ (for $t = 0, 1, 2, \ldots$) obtained on the top from a single run of the IBM, and the bottom by summing the daily number of cases for 500 IMB runs. On the right-hand side, we apply formula (6.5) (with $I_0 = 1000$) to the daily number of cases obtained from the IBM. The top two figures correspond to $I_0 = 1000$, and the bottom two figures to $I_0 = 500 \times 1000$.

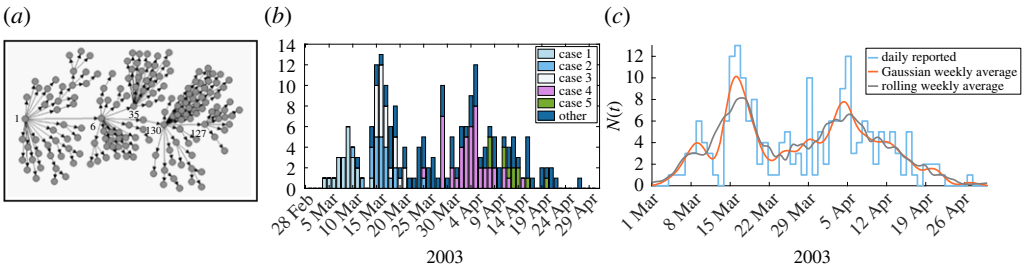

Figure 7. (*a*) We plot the contact network of the five super spreader cases in the SARS epidemic in Singapore in 2003 [21]. The super spreaders are patient 1, patient 6, patient 35, patient 130 and patient 127. (*b*) Daily reported cases from Singapore for the epidemic of SARS in 2003. Case 1 generated 21 cases and 3 suspected cases, case 2 generated 23 cases and 5 suspected cases, case 3 generated 23 cases and 18 suspected cases, case 4 generated 40 cases and 22 suspected cases, case 5 generated 15 cases and 0 suspected cases [21]. The cases 1,2,3,4,5 correspond, respectively, to patients 1, 6, 35, 130 and 127. (*c*) Regularizations of the daily cases data from the SARS-CoV-1 outbreak in Singapore [21]. The blue curve corresponds to a step function, the orange curve to a Gaussian weekly average, and the grey curve to a rolling weekly average. The applications in figure 8 are done with the 'Rolling Weekly' regularization.

Figure 7*a*,*b* presents the time series of reported cases by source of infection and date of fever onset and (*c*) presents three representations of these data in continuous time: as a step function, regularized by Gaussian average and rolling weekly average. In figure 8, we apply the continuous-time model to the rolling weekly regularization of the data. Similar to the reconstruction of $R_0(a)$ presented in figures 4–6, the basic reproduction number $R_0(a)$ becomes negative after a given age. Our interpretation is that the data are far from perfect and involve sampling errors and probably a large number of undetected cases.

The fact that the transmission rate is subject to variations in time could also explain this negativity. In figure 8, we apply the discrete model (1.6) to the original data for different values of I_0. In figure S10 in the electronic supplementary material, we transform the data by taking

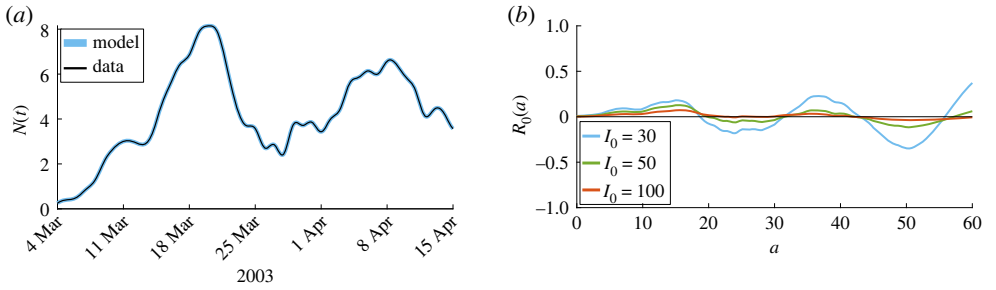

Figure 8. (*a*) Regularized data of the SARS-CoV-1 outbreak in Singapore in 2003 [21] (black line) and the numerical solution of the model (1.1) with $I_0 = 30$ and $R_0(a)$ computed by (1.6) (blue line). The solutions $N(t)$ of the model (1.1) with $I_0 = 50$ and $I_0 = 100$ are exactly the same when we use the corresponding $R_0(a)$, therefore, they are not represented here. (*b*) Numerical solution of the $R_0(a)$ function computed by using the continuous model (1.4) with $I_0 = 30$, $I_0 = 50$ and $I_0 = 100$.

advantage of the information on the source of infection given in [21]. We fix an incubation period of 5 days, which corresponds to the average incubation period reported in [21]. Then we shift all secondary cases produced by the six sources identified in the article [21] to the same origin, as if all cases had been produced by the same cluster of six individuals. We present the data on the left-hand side of figure S11 of the electronic supplementary material, and apply the method to obtain $R_0(a)$ for parameters $I_0 = 30$, $I_0 = 50$, $I_0 = 100$.

In this work, we did not consider the problem of uncertainty in observed data. Nevertheless, this uncertainty could explain the negative values of the solutions observed on the right-hand side of figure 8.

9. Discussion

We see from the numerical simulations in §7 that the initial number of infected I_0 has a very significant influence on the value of the basic daily reproduction numbers $R_0(d)$: these decrease sharply with I_0, until they become negative and their fluctuations increase in stochastic simulations. This tendency to negativity for small I_0 and these fluctuations in the stochastic case are only corrected when the results are averaged for a large number of stochastic simulations (500). It can also be noted that the stochastic simulations lead to a behaviour of the hyper exponential type in the coefficient of variation of the secondary cases produced by an infectious individual, that is to say that it is relatively constant and much greater than 1. This phenomenon is to be related to the exponential character of the gamma distribution used in the simulations.

In stochastic simulations, we observe the same behaviour for the different curves related to the $R_0(a)$ curves sample, but the expectation of this curves sample considerably attenuates these fluctuations and the coefficient of variation of the curves remains approximately constant, while being greater than 1, as in the case of hyperexponential distributions, in agreement with the exponential character of part D of equation (1.7) defining $R_0(a)$.

Concerning the clusters, from observations made during investigations of the start of the outbreak in some countries [22,–35], it is possible to get spatial and temporal information on the start of the epidemic, but these studies rarely allow the estimation of the parameters S_0 and τ in the concerned population and worse, they give no indication of how long they remain constant. Here, we assumed that they remained constant only during the period of exponential growth of new cases observed.

Our work provides a method to reconstruct the daily basic reproduction number $R_0(a)$ from the daily reported cases data, as long as we consider a cluster starting from a single infected. This is a strong assumption which is usually neglected. It is extremely hard to find in the literature a dataset which satisfies this assumption. For COVID-19, we did not find any publication including

suitable data. While not published yet, we believe that this kind of data could be gathered by a detailed contact-tracing and—duly anonymized—could be made available by request. That would allow the future development of more realistic and accurate methods for the analysis and forecasting of epidemics.

Data accessibility. The data are provided in electronic supplementary material [36].

Authors' contributions. J.D.: conceptualization, methodology, writing—original draft, writing—review and editing; Q.G.: conceptualization, methodology, software, writing—original draft, writing—review and editing; Y.M.: conceptualization, methodology, writing—original draft, writing—review and editing; P.M.: conceptualization, methodology, software, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. No funding has been received for this article.

References

1. Alvarez L, Colom M, Morel JD, Morel JM. 2021 Computing the daily reproduction number of COVID-19 by inverting the renewal equation using a variational technique. *Proc. Natl Acad. Sci. USA* **118**, e2105112118. (doi:10.1073/pnas.2105112118)
2. Alvarez L, Morel J-D, Morel J-M. 2022 Modeling COVID-19 incidence by the renewal equation after removal of administrative bias and noise. *Biology* **11**, 540. (doi:10.3390/biology11040540)
3. Nishiura H, Chowell G. 2009 The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In *Mathematical and Statistical Estimation Approaches in Epidemiology* (eds G Chowell, JM Hyman, LMA Bettencourt, C Castillo-Chavez), pp. 103–121. Dordrecht, Netherlands: Springer.
4. Bernoulli D. 1760 Essai d'une nouvelle analyse de la mortalité causé par la petite vérole et des avantages de l'inoculation pour la prévenir. *Mém. Math. Phys. Acad. R. Sci. Paris*, 1–45.
5. Demongeot J, Oshinubi K, Rachdi M, Seligmann H, Thuderoz F, Waku J. 2021 Estimation of daily reproduction rates in COVID-19 outbreak. *Computation* **9**, 109. (doi:10.3390/computation9100109)
6. Worobey M *et al.* 2016 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* **539**, 98–101. (doi:10.1038/nature19827)
7. Demongeot J, Griette Q, Magal P. 2020 SI epidemic model applied to COVID-19 data in mainland China. *R. Soc. Open Sci.* **7**, 201878. (doi:10.1098/rsos.201878)
8. Demongeot J, Griette Q, Magal P, Webb G. 2022 Modeling vaccine efficacy for COVID-19 outbreak in New York city. *Biology* **11**, 345. (doi:10.3390/biology11030345)
9. Griette Q, Demongeot J, Magal P. 2021 A robust phenomenological approach to investigate COVID-19 data for France. *Math. Appl. Sci. Eng.* **2**, 149–218. (doi:10.1101/2021.02.10.21251500)
10. Griette Q, Demongeot J, Magal P. 2022 What can we learn from COVID-19 data by using epidemic models with unidentified infectious cases? *Math. Biosci. Eng.* **19**, 537–594. (doi:10.3934/mbe.2022025)
11. Liu Z, Magal P, Seydi O, Webb G. 2020 Understanding unreported cases in the COVID-19 epidemic outbreak in Wuhan, China, and the importance of major public health interventions. *Biology* **9**, 50. (doi:10.3390/biology9030050)
12. Clément F, Laroche B, Robin F. 2019 Analysis and numerical simulation of an inverse problem for a structured cell population dynamics model. *Math. Biosci. Eng.* **16**, 3018–3046. (doi:10.3934/mbe.2019150)
13. Gyllenberg M, Osipov A, Päivärinta L. 2002 The inverse problem of linear age-structured population dynamics. *J. Evol. Equ.* **2**, 223–239. (doi:10.1007/s00028-002-8087-9)
14. Krivtsov V, Yevkin O. 2013 Estimation of G-renewal process parameters as an ill-posed inverse problem. *Reliab. Eng. Syst. Saf.* **115**, 10–18. (doi:10.1016/j.ress.2013.02.005)
15. Pijpers FP. 2021 A non-parametric method for determining epidemiological reproduction numbers. *J. Math. Biol.* **82**, 1–21. (doi:10.1007/s00285-021-01590-6)
16. Kermack WO, McKendrick AG. 1932 Contributions to the mathematical theory of epidemics: II. *Proc. R. Soc. Lond. B* **138**, 55–83. (doi:10.1098/rspa.1932.0171)
17. Bakhta A, Boiveau T, Maday Y, Mula O. 2020 Epidemiological forecasting with model reduction of compartmental models. application to the COVID-19 pandemic. *Biology* **10**, 22. (doi:10.3390/biology10010022)

18. Ducrot A, Magal P. In preparation. *A semigroup approach for Volterra integral equation of convolution type*.
19. Chao DL, Halloran ME, Obenchain Jr VJ, Longini IM. 2010 FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Comput. Biol.* **6**, e1000656. (doi:10.1371/journal.pcbi.1000656)
20. Pan Y, Zhang D, Yang P, Poon LLM, Wang Q. 2020 Viral load of SARS-CoV-2 in clinical samples. *Lancet Infect. Dis.* **20**, 411–412. (doi:10.1016/S1473-3099(20)30113-4)
21. Centers for Disease Control and Prevention (CDC). 2005 Severe acute respiratory syndrome Singapore. *MMWR. Morbidity and mortality weekly report*, **52**. www.cdc.gov/mmwr/preview/mmwrhtml/mm5218a1.htm.
22. Adam DC, Wu P, Wong JY, Lau EH, Tsang TK, Cauchemez S, Leung GM, Cowling BJ. 2020 Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.* **26**, 1714–1719. (doi:10.1038/s41591-020-1092-0)
23. Böhmer MM *et al.* 2020 Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. *Lancet Infect. Dis.* **20**, 920–928. (doi:10.1016/S1473-3099(20)30314-5)
24. Chan TC, King CC. 2011 Surveillance and epidemiology of infectious diseases using spatial and temporal lustering methods. In *Infectious disease informatics and biosurveillance* (eds C Castillo-Chavez, H Chen, WB Lober, M Thurmond, D Zeng), pp. 207–234. Springer, Boston, MA.
25. Desjardins MR, Hohl A, Delmelle EM. 2020 Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: detecting and evaluating emerging clusters. *Appl. Geogr.* **118**, 102202. (doi:10.1016/j.apgeog.2020.102202)
26. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, Hens N. 2020 Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance* **25**, 2000257. (doi:10.2807/1560-7917.ES.2020.25.17.2000257)
27. Guttmann A, Ouchchane L, Li X, Perthus I, Gaudart J, Demongeot J, Boire JY. 2013 Performance map of a cluster detection test using extended power. *Int. J. Health Geogr.* **12**, 1–10. (doi:10.1186/1476-072X-12-47)
28. Han L *et al.* 2021 Exploring the clinical characteristics of COVID-19 clusters identified using factor analysis of mixed data-based cluster analysis. *Front. Med.* **8**, 644724. (doi:10.3389/fmed.2021.644724)
29. Hisada S, Murayama T, Tsubouchi K, Fujita S, Yada S, Wakamiya S, Aramaki E. 2020 Surveillance of early stage COVID-19 clusters using search query logs and mobile device-based location information. *Sci. Rep.* **10**, 1–8. (doi:10.1038/s41598-020-75771-6)
30. Jing QL *et al.* 2020 Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 1141–1150. (doi:10.1016/S1473-3099(20)30471-0)
31. Ladoy A, Opota O, Carron PN, Guessous I, Vuilleumier S, Joost S, Greub G. 2021 Size and duration of COVID-19 clusters go along with a high SARS-CoV-2 viral load: a spatio-temporal investigation in Vaud state, Switzerland. *Sci. Total Environ.* **787**, 147483. (doi:10.1016/j.scitotenv.2021.147483)
32. Pung R *et al.* 2020 Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *Lancet* **395**, 1039–1046. (doi:10.1016/S0140-6736(20)30528-6)
33. Shams F, Abbas A, Khan W, Khan US, Nawaz R. 2022 A death, infection, and recovery (DIR) model to forecast the COVID-19 spread. *Comput. Methods Programs Biomed. Update* **2**, 100047. (doi:10.1016/j.cmpbup.2021.100047)
34. Tariq A, Lee Y, Roosa K, Blumberg S, Yan P, Ma S, Chowell G. 2020 Real-time monitoring the transmission potential of COVID-19 in Singapore, March 2020. *BMC Med.* **18**, 1–14. (doi:10.1186/s12916-020-01615-9)
35. Yong SEF *et al.* 2020 Connecting clusters of COVID-19: an epidemiological and serological investigation. *Lancet Infect. Dis.* **20**, 809–815. (doi:10.1016/S1473-3099(20)30273-5)
36. Demongeot J, Griette Q, Maday Y, Magal P. 2023 A Kermack–McKendrick model with age of infection starting from a single or multiple cohorts of infected patients. Figshare. (doi:10.6084/m9.figshare.c.6500603)