Journal of Mathematical Biology (2025) 91:19
https://doi.org/10.1007/s00285-025-02249-2

On final and peak sizes of an epidemic with latency and effect of behaviour change

Tianyu Cheng[1] · **Xingfu Zou**[1]

Received: 30 October 2024 / Revised: 30 April 2025 / Accepted: 26 June 2025

Abstract
In this paper, we use the renewal equation approach to explore the impact of behaviour change and/or non-pharmaceutical interventions (NPIs) on the final size and peak size of an infectious disease without demography. To this end, we derive the renewal equations (REs) for the force of infection (both instantaneous and cumulative) that have reflected the NPIs and/or behaviour change by the notion of practically susceptible population. A novelty in these REs is that they contain time-varying kernels arising from the incorporation of effect of behaviour change. We then build the new REs into the Kermack–McKendrick model to obtain a general full model. Following Breda et al. (J Biol Dyn 6(sup2):103–117, 2012) and Diekmann et al. (Proc Natl Acad Sci USA 118(39):e2106332118, 2021), we analyze this new model to derive a general formula for the final size relation, by which we find that the final size relation depends not only on the basic reproduction number $\mathcal{R}_0$ but also on other associated values that reflect the impact of behaviour change. Specifically, we demonstrate that behaviour change can reduce the infection peak and herd immunity threshold in some specific models.

Keywords Infection-age · Behaviour change · Non-pharmaceutical intervention · Practically susceptible · Final size · Renewal equation · Force-of-infection

Mathematics Subject Classification 34D20 · 34K18 · 34K20 · 91A22 · 92B20 · 92D30

Partially supported by NSERC of Canada (RGPIN-2022-04744).

✉ Xingfu Zou
xzou@uwo.ca

[1] Department of Mathematics, University of Western Ontario, London, ON N6A 5B7, Canada

Published online: 22 July 2025

1 Introduction

Many new emerging diseases have much shorter durations than the life span of the human host. Thus, in modelling the dynamics of such diseases, one typically ignores the demography of the host and just focuses on the transmission dynamics. For such a model, the disease eventually dies out, and hence, the long-term disease dynamics are no longer a major concern. Instead, the scale of the epidemic is generally the major concern, which can be reflected by the final size and the peak size. Taking the classic Kermack–McKendrick (K–M) disease model in Kermack and McKendrick (1927) as an example, the final size r_∞ is determined in Kermack and McKendrick (1927) by the equation

$$r_\infty = 1 - e^{-\mathcal{R}_0 r_\infty} \tag{1.1}$$

where $r_\infty = \lim_{t\to\infty} r(t)$ with $r(t)$ representing the fraction of recovered host population in a closed population at time t.

Similar formulas have been obtained for final sizes of epidemics in some other circumstances (some expressed *in terms of the susceptible population* rather than the recovered population). For example, the simplest case of the K–M model in Kermack and McKendrick (1927) is described by the following ODE system

$$\begin{cases} S'(t) = -\beta I(t)S(t), \\ I'(t) = \beta I(t)S(t) - \gamma I(t), \\ R'(t) = \gamma I(t), \end{cases} \tag{1.2}$$

for which the final size equation can actually be easily obtained in terms of $S(\infty)$ by analyzing the trajectories of the $I - S$ equations and using the relation $S(\infty) + R(\infty) = S(0) =: S_0$. See, e.g., Miller (2012)). Along the same line, Magal et al. (2016) investigated the final size of the *two patch version* of (1.2); while Ma and Earn (2006) explored the final size of *multi-stage version* of (1.2). There are some other more works that discuss the final size of various generalizations of the epidemic model (1.2) by analyzing the *S-I* equations, e.g., Anderson and Watson (1980), Arino et al. (2007), Andreasen (2011), Brauer (2008), Ketcheson (2021), Feng (2007), Kröger et al. (2021) and the references therein. We point out that this method is valid in these models mainly because the FOI in these models is linear in I (i.e. βI) corresponding to the mass action incidence rate. This can be clearly reflected in Arino et al. (2007): while a general FOI function is used for other topics on disease dynamics, however, when it comes to the *final size* of the epidemic version of the model, the authors have to restrict themselves to the linear FOI.

On the other hand, note that during an epidemic, especially a severe epidemic (e.g., the recent covid-19 pandemic), due to the host's fear of the disease and/or the non-pharmaceutical interventions (NPIs), some host individuals will behave more pre-cautiously and be more restrictive in their social activities. This implies that only a fraction P of the *epidemically susceptible* host is *practically susceptible*: $S_p(t) = PS(t)$. Here the fraction $P \in [0, 1]$ reflects the precaution level reflecting the effect of behaviour change of the public, and it typically depends on the severity

or prevalence of the epidemic, letting it be measured by $L(t)$ and then $P = P(L(t))$ changes as the epidemic evolves. By its background, $P(L)$ satisfies the following conditions:

$$\frac{dP(L)}{dL} \leq 0 \quad \text{with} \quad P(0) \leq 1 \quad \text{and} \quad \lim_{L\to\infty} P(L) \geq 0, \tag{1.3}$$

meaning that the more severe the epidemic is, the more cautious the public is. The severity level function L can be described by some disease-related variables, e.g., the infected population $I(t)$ or the weighted sum of these disease variables, and it contains information on disease severity. Possible candidates include

$$L(t) = \sum_{i=0}^{n} w_i I(t-\tau_i) \quad \text{and} \quad L(t) = \int_0^{\tau} w(\xi) I(t-\xi)\, d\xi. \tag{1.4}$$

where $w_i,\ i = 0, 1, \ldots n$ and $w(\xi),\ \xi \in [0, \tau]$ are weights for the past values of $I(t)$. See Cheng and Zou (2022) for some details.

For convenience of notation and calculations in later places, we write $P(L(t))$ as $P_l(t)$ to mean that the fraction P is time varying through the severity level. Replacing $S(t)$ by $S_p(t)$ in the mass action incidence rate $\beta I(t)S(t)$ and standard incidence rate $\beta I(t)S(t)/[I(t) + S(t)]$ respectively, we obtain the respective new incidence rates

$$\beta I(t)S_p(t) = [\beta P_l(t)I(t)] \cdot S(t) \ \text{ and } \ \frac{\beta I(t)S_p(t)}{I(t)+S_p(t)} = \left[\frac{\beta P_l(t)I(t)}{I(t)+P_l(t)S(t)}\right] \cdot S(t),$$

which leads to the respective new FOIs given by

$$F(t) = \beta P_l(t)I(t) \ \text{ and } \ F(t) = \frac{\beta P_l(t)I(t)}{I(t)+P_l(t)S(t)}. \tag{1.5}$$

Such new FOIs are no longer linear in I and depend on the past of $I(t)$. In Cheng and Zou (2022) and a follow-up work (Cheng and Zou 2024), using the revised FOIs that reflect the effects of behaviour change as explained above, we explored the impact of such effects on the *long-term disease dynamics* for some disease models *with demography*. Now, as we mentioned in the beginning, our interest in this work is in the final size and peak size of an epidemic that is deemed to die out (due to the nature of "without demography"). Then there arises the question:

(Q1) What would be the impact of the host's response/precaution reflected by $P(L)$) on the final size and peak size of a disease model without demography?

Notice the above revised FOI $\beta P(L(t))I(t)$ is <u>no longer linear</u> in I (noting that $L(t)$ depends on $I(t)$) and could be of various forms involving $I(t)$ and its past values (see Cheng and Zou 2022, 2024). As such, it becomes very challenging, if not impossible, to derive the final size equation by analyzing the trajectories of the (1.2) with βI replaced by $\beta P(L(t))I(t)$. And if one wishes to also incorporate disease latency

leading to a model with delays, the final size becomes even more challenging. These drive us to consider an alternative approach.

Notice that Breda et al. (2012) reformulated the *general K–M mode* in Kermack and McKendrick (1927) in terms of the RE for the force-of-infection (FOI) $F(t)$

$$F(t) = \int_0^\infty F(t-\tau)S(t-\tau)A(\tau)d\tau, \tag{1.6}$$

and the cumulative force-of-infection (CFOI) $y(t)$

$$y(t) = \int_0^\infty (1 - e^{-y(t-\tau)})S(-\infty)A(\tau)d\tau. \tag{1.7}$$

Here $A(\tau)$ describes the expected contribution of the infected population to the force-of-infection. Indeed, by the RE approach, the authors established the following equation for $y(\infty)$

$$y(\infty) = \mathcal{R}_0(1 - e^{-y(\infty)}), \tag{1.8}$$

and identified the relation

$$r_\infty = \frac{y(\infty)}{\mathcal{R}_0} \tag{1.9}$$

which transforms the final size equation (1.8) for the CFOI to the final size equation (1.1) for the recovered fraction. For the special case (1.2), Brauer (2008) also used RE for the FOI to reproduce the final size equation.

The above brief illustration of the RE approach in terms of FOI shows that FOI is indeed an important notion and offers an alternative candidate for employing the RE approach. This makes us wonder that

(Q2) with the above revised FOI $F(t) = \beta P(L(t))I(t)$ mediated by the precaution of the public that reflects the behaviour change due to host's fear of the disease can we apply the RE approach to address **(Q1)**?

It is well-known that most compartmental epidemic models can be described by REs, and hence, we wish to employ the RE theory to address these issues. Biologically, REs give a reasonable way to trace renewals/reproductions of the given class/population over time. The idea was initially developed by Lotka (1907), Sharpe et al. (1911) and used by many researchers, e.g., Feller (1940, 1941), Inaba (2017), Lebreton (1996), in modelling demography and species reproduction. Particularly, Diekmann et al. re-drew researchers' attention to the RE method and demonstrated its powerfulness/usefulness in mathematical epidemiology-modelling in several works (e.g., Breda et al. 2012; Champredon et al. 2018; Diekmann and Heesterbeek 2000; Diekmann et al. 2021; Diekmann and Inaba 2023).

Encouraged by these works, in this paper, we will tackle the above two questions by the RE approach. To this end, in Sect. 2, we will first present some preliminaries that involve infection age. Going through the references (Breda et al. 2012; Ker-

mack and McKendrick 1927) intensively, we find that *age/stage density* is the key to associating REs and compartment systems—age density deduces a RE and a compartment system can be obtained through the age density. In Sect. 3, we derive the renewal equations for the revised FOI function and the corresponding incidence rate function. Then, in Sect. 4, we use the REs for the revised FOI and the incidence rate to derive equations that govern the final size. Section 5 is dedicated to some special cases of the general model represented by the REs model in Sect. 3, for which we are able to obtain more detailed and more explicit results, particularly some results on the peak size, which is, in general, very challenging (if not impossible) for the general case. We conclude the paper with a discussion section (Sect. 6), in which we discuss the implications of our results and some possible extensions of this work.

We point out that our new REs contain time-varying kernels due to the behaviour change. It is the *time varying nature* of the kernel that makes our model more novel, the analysis more challenging, and the results more interesting both mathematically and epidemiologically. Our new model provides a general and plausible framework for estimating the final epidemic size and even peak size (in some special situations) when considering NPIs and/or behaviour changes during an epidemic.

2 Preliminaries on infection age structure and related notions

2.1 Density of infected class with respect to infection age

Denote by $a \in [0, \infty)$ the infection age, which defines the length of time an individual has been infected, and let $u(t, a)$ be the density (with respect to a) of the infected class at time t with infection age a. Let $\gamma(a)$ denote the recovery rate of the infected population with age a and assume that $\gamma(\infty) < \infty$. For the age density of infected individuals, there is the well-known McKendrick/von Foerster equation:

$$\frac{\partial u(t,a)}{\partial a} + \frac{\partial u(t,a)}{\partial t} = -\gamma(a)u(t,a). \tag{2.1}$$

Firstly, we note that the characteristics of (2.1) define a family of lines with equation $t = t_0 + a$, where $t_0 > 0$ is considered the "birthdate" (infection time) of an infected individual. Then, the infection age density along the characteristic is $u(t, a) = u(t_0 + a, a) := u_{t_0}(a)$ which solves

$$\frac{du_{t_0}(a)}{da} = -\gamma(a)u_{t_0}(a). \tag{2.2}$$

The solution of (2.2) for $t > a$ then is:

$$u(t,a) = u_{t_0}(a) = u_{t_0}(0)e^{-\int_0^a \gamma(x)dx} = u(t_0,0)e^{-\int_0^a \gamma(x)dx} = u(t-a,0)e^{-\int_0^a \gamma(x)dx}. \tag{2.3}$$

Let

$$\sigma(a) := e^{-\int_0^a \gamma(x)dx},$$

which expresses the probability of a newly infected individual remaining infected at infection age a. That means a fraction $\sigma(a)$ of the population infected at time $t-a$ will still remain infected at time t with age a. Simple calculation yields

$$\sigma'(a) = -\gamma(a)\sigma(a) < 0$$

Naturally, the following assumption on $\sigma(a)$

$$\sigma(\infty) = e^{-\int_0^\infty \gamma(x)dx} = 0, \tag{2.4}$$

should be imposed since no individual will "survive" to infinity age. Accordingly $u(t,\infty) = 0$ for any $t \in \mathbb{R}$.

Next, we need to track those infected individuals who are already infected at $t = 0$ with age a_0 and still remain infected at t with age $a = a_0 + t$. Let $u_0(a_0) = u(0, a_0)$, which is the initial age distribution of the infected class. Solving (2.1) for $a > t$, one obtains

$$u(t,a) = u_0(a-t)e^{-\int_{a-t}^a \gamma(x)dx} = u_0(a-t)\frac{\sigma(a)}{\sigma(a-t)} \tag{2.5}$$

Here, $\sigma(a)/\sigma(a-t)$ describes the probability of those already infected initially at $t = 0$ with age $a_0 = a - t$ still remaining in the infected class at time t (with infection age a).

Summarizing the above, the solution of (2.1) is expressed piecewise as below

$$u(t,a) = \begin{cases} u(t-a,0)\sigma(a), & a < t \\ u_0(a-t)\frac{\sigma(a)}{\sigma(a-t)}, & a > t. \end{cases} \tag{2.6}$$

The piecewise nature accounts for the two sources for $u(t, a)$: one is from the initially infected individuals for $a > t$ represented by $u_0(a-t) = u(0, a-t)$, and the other is from the new infections ("birth" or "production") during $[0, t-a]$ for $a < t$ represented by $B(t-a) = u(t-a, 0)$.

Many diseases have latency, and the latent period can vary from individual to individual. For some diseases, the individual latencies vary slightly; for such a disease, one may assume a fixed constant latency. If considering a fixed latency $\tau \geq 0$,

$$E(t) := \int_0^\tau u(t,a)da, \quad I(t) := \int_\tau^\infty u(t,a)da,$$

then represent the total numbers of latent class and infectious class, respectively. Accordingly, $U(t) := E(t) + I(t)$ is the total number of infected individuals and

$$R(t) := \int_0^\infty \gamma(a)u(t,a)da,$$

sums up all those who have recovered at different infection ages and, hence, gives the total number of recovered classes at time *t*.

Throughout this work, we use the following classical/standard notations:

$$s(t) := S(t)/N_0,\ \ s(\infty) := s_\infty \quad \text{and} \quad r(t) := R(t)/N_0,\ \ r(\infty) := r_\infty. \tag{2.7}$$

2.2 The related mean period

Consider the recovery age as a random variable ξ. Then

$$\mathbb{P}(0 < \xi \le a) = 1 - \sigma(a) = \int_0^a \gamma(x)\sigma(x)dx$$

Accordingly, the probability density function of ξ is $f(a) = \gamma(a)\sigma(a)$. Assuming fixed constant latency $\tau > 0$ as stated above, then

$$\begin{aligned} T_E &= \frac{\int_0^\tau x\gamma(x)\sigma(x)dx}{\int_0^\tau \gamma(x)\sigma(x)dx} = \frac{\int_0^\tau \sigma(x)dx}{1-\sigma(\tau)} - \frac{\tau\sigma(\tau)}{1-\sigma(\tau)} \quad \text{and} \\ T_I &= \frac{\int_\tau^{+\infty} x\gamma(x)\sigma(x)dx}{\int_\tau^\infty \gamma(x)\sigma(x)dx} = \frac{\int_\tau^\infty \sigma(x)dx}{\sigma(\tau)} + \tau \end{aligned} \tag{2.8}$$

represent the mean time infected individuals stay in the latent class and the mean time of an infected individual stays in the infectious class, respectively. Consequently, $T_I - \tau$ is the mean infectious period and

$$\begin{aligned} T_{tol} &= \int_0^\infty x\gamma(x)\sigma(x)dx = -\int_0^\infty xd\sigma(x) = \int_0^\infty \sigma(x)dx - x\sigma(x)|_{x=0}^{x=\infty} \\ &= \int_0^\infty \sigma(x)dx \\ &= \int_0^\tau \sigma(x)dx + \int_\tau^\infty \sigma(x)dx \\ &= \sigma(\tau)T_I + (1-\sigma(\tau))T_E \end{aligned} \tag{2.9}$$

is the mean infected period, which is the weighted sum of T_E and T_I.

3 Derivation of the model

3.1 The derivation of the REs for $B(t)$ and FOI $F(t)$

After the preparations in Sect. 2, we are now in a position to follow the idea in Breda et al. (2012) to formulate our new REs in terms of the new FOIs in (1.5). Assume that the infectivity of infected individuals with infection age a is $\beta(a) \in [0, 1]$ for $a \in [\tau, \infty)$. Then the total infectivity of the infectious population at time t is

$$C(t) := \int_{\tau}^{\infty} \beta(a)u(t, a)da. \tag{3.1}$$

As is known, this is traditionally defined as the force-of-infection when focusing on physiological infectivity (see, e.g., Anderson and May 1991). Particularly, if $\beta(a)$ is independent on a, $C(t) = \beta I(t)$.

Now, with the new FOI given in (1.5), the incidence rate is accordingly revised to

$$\begin{aligned} B(t) &:= \int_{\tau}^{\infty} [\beta(a)P_l(t)S(t)u(t, a)]da = S_P(t)C(t) \\ &= P_l(t)S(t)C(t) = [P_l(t)C(t)] \cdot S(t); \end{aligned} \tag{3.2}$$

accordingly, the force-of-infection in (3.1) is revised to

$$F(t) := P_l(t)C(t) \tag{3.3}$$

In (3.3), $C(t)$ and $P(L)$ measure the efficiency of physiological infection (focusing more on features of epidemics) and effect of behaviour change respectively. In other words, F is a result of combining the infectivity and effect of behaviour change.

Obviously, $u(t, 0) = B(t)$ holds, since the incidence accounts for new infections occurring per unit time. Moreover, for $t > \tau$, the incidence rate $B(t)$ can also be expressed as

$$\begin{aligned} B(t) &= \int_{\tau}^{\infty} \beta(a)S_P(t)u(t, a)da \\ &= S_P(t)\left(\int_{\tau}^{t} \beta(a)u(t, a)da + \int_{t}^{\infty} \beta(a)u(t, a)da\right) \\ &= S_P(t)\left(\int_{\tau}^{t} \varphi(a)B(t - a)da + G(t)\right) \end{aligned} \tag{3.4}$$

where

$$\varphi(a) := \begin{cases} 0, & a \in [0, \tau) \\ \beta(a)\sigma(a), & a \in [\tau, \infty) \end{cases} \tag{3.5}$$

and

$$G(t) := \int_t^\infty \beta(a)u_0(a-t)\frac{\sigma(a)}{\sigma(a-t)}da = \int_t^\infty u_0(a-t)\frac{\varphi(a)}{\sigma(a-t)}da = \int_0^\infty u_0(\eta)\frac{\varphi(\eta+t)}{\sigma(\eta)}d\eta. \tag{3.6}$$

Here $\varphi(a)$ is continuous in $a \in [\tau, \infty)$, and it is the expected contribution to the FOI from those infected individuals with infection age a (according to Breda et al. 2012). Note that $T_{tol} < \infty$ ensures $\int_\tau^\infty \varphi(a)da < \infty$.

When $0 \le t < \tau$, $a > \tau$ implies $a > t$ and hence

$$B_\tau(t) = S_P(t)C(t) = S_P(t)\int_\tau^\infty \beta(a)u(t,a)da = S_P(t)G_\tau(t) \tag{3.7}$$

where

$$G_\tau(t) = \int_\tau^\infty \beta(a)u(t,a)da = \int_\tau^\infty \beta(a)u_0(a-t)\frac{\sigma(a)}{\sigma(a-t)}da = \int_\tau^\infty u_0(a-t)\frac{\varphi(a)}{\sigma(a-t)}da = \int_{\tau-t}^\infty u_0(\eta)\frac{\varphi(\eta+t)}{\sigma(\eta)}d\eta. \tag{3.8}$$

Denote

$$G^\diamond(t) = \begin{cases} G(t) & t \ge \tau, \\ G_\tau(t) & t < \tau. \end{cases} \tag{3.9}$$

$G^\diamond(t)$ reflects the contribution to new infection at t by those already infective at $t = 0$. $G^\diamond(t)$ is continuous at $t = \tau$. Because $\varphi(a)$ goes to zero when the infection age a of an individual tends to infinity, $G(t)$ also tends to zero, i.e. $\lim_{t\to\infty} G^\diamond(t) = \lim_{t\to\infty} G(t) = 0$.

Following Breda et al. (2012), Diekmann et al. (2021), we may extend the domain of $B(t)$ to include negative axis, so that we actually have

$$G(t) = \int_t^\infty \varphi(a)B(t-a)da \quad \text{and} \quad G_\tau(t) = \int_\tau^\infty \varphi(a)B(t-a)da,$$

due to the fact that $u_0(a-t) = B(t-a)\sigma(a-t)$. Combining the above, we then have obtained the REs for $B(t)$ and $F(t)$ as

$$B(t) = S_P(t)\int_\tau^\infty \varphi(a)B(t-a)da \tag{3.10}$$

and

$$F(t) = P_l(t)\int_\tau^\infty \varphi(a)B(t-a)da = \int_\tau^\infty \underbrace{P_l(t)\phi(a)}_{\text{Revised kernel}} F(t-a)S(t-a)\,da \tag{3.11}$$

for $t \in (-\infty, \infty)$ rather than for $t \in (0, \infty)$ only. This allows us to consider the initial time to be at $-\infty$ for both (3.10) and (3.11), as is done in Breda et al. (2012), Diekmann et al. (2021). We point out that in contrast to (1.6), the RE (3.11) is no longer a traditional scalar RE because the kernel $P_l(t)\phi(a) = P(L(t))\phi(a)$ depends not only on the infection age a but also on the time variable t through the severity $L(t)$. Moreover, the severity measure may involve the past values of $I(t)$, as demonstrated in (1.4). This means that even for the current value $B(t)$ or $F(t)$, we need knowledge/information of not only the current *"practically susceptible population"*, but also the entire history of B over the interval $(-\infty, t]$.

We point out that the REs (3.10) and (3.11) naturally lead to the iterated transition property (semi-group property) and allow us to track how the current value depends on the past values. For example, for any $a_1, a_2 \geq 0$, from (3.10), we can obtain

$$B(t) = S_P(t)\int_\tau^\infty \varphi(a_2)\left\{S_P(t-a_2)\int_\tau^\infty \varphi(a_1)B(t-a_2-a_1)da_1\right\}da_2 \tag{3.12}$$

which is a result of two iterative trackings. See Fig. 1 for an illustration of (3.12).

3.2 Full general model

With the above preparation, a general model without demography can be formulated by RE as

$$\begin{cases} S'(t) = -B(t) = -F(t)S(t) \\ F(t) = P_l(t)\int_\tau^\infty \varphi(a)F(t-a)S(t-a)da \end{cases} \tag{3.13}$$

Corresponding to the model (3.13) with $\varphi(a)$ given in (3.5), we have the following compartmental model (see details in Appendix):

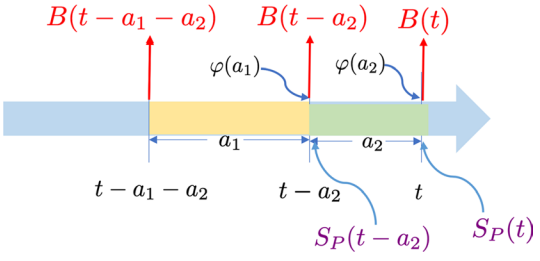

Fig. 1 Depiction of the RE (3.12) of $B(t)$

$$\begin{cases} S'(t) = -F(t)S(t) \\ E'(t) = -\int_0^\tau \gamma(a)u(t,a)da - u(t,\tau) + F(t)S(t) \\ I'(t) = -\int_\tau^\infty \gamma(a)u(t,a)da + u(t,\tau) \\ R'(t) = \int_0^\infty \gamma(a)u(t,a)da \end{cases} \tag{3.14}$$

We assume there is no recovery individual at the epidemic's beginning time $t = t_0$ (i.e., $R(t_0) = 0$) throughout this paper. In the sequel, as we proceed further, we will take $t_0 = -\infty$ or $t_0 = 0$, depending on the purpose.

Remark 3.1 It is worth noting that (3.13) and (3.14) can both be obtained from the equation (2.1), respectively. Age density of the infected population $u(t, a)$ establishes a bridge between the REs model (3.13) and the compartmental model (3.14).

Observe that in (3.14), the equations of E' and I' couple via the term $u(t, \tau)$, which is the rate at which The infectious class gains from the latent class. Now we determine $u(t, \tau)$ by (2.6) as below.

For $t \geq \tau$ (long-time), by (2.6),

$$u(t,\tau) = u(t-\tau,0)\sigma(\tau) = B(t-\tau)\sigma(\tau).$$

That means that, with probability $\sigma(\tau)$, those individuals who get infected at time $t - \tau$ will enter the infectious class from the latent class at time t.

If $t < \tau$ (short-time), again by (2.6),

$$u(t,\tau) = u_0(\tau - t)\frac{\sigma(\tau)}{\sigma(\tau - t)}.$$

At such a time t, only those who are already infected at time $t = 0$ are possible to enter the infectious class (with probability $\sigma(\tau)/\sigma(\tau - t)$), because at time t, all newly infected individuals during $[0, t]$ have no chance to become infectious (their infection ages are all less than the latency τ. Moreover, noting that

$$u_0(\tau - t) = u(-(\tau - t), 0)\sigma(\tau - t) = B(t - \tau)\sigma(\tau - t),$$

we have actually also have $u(t, \tau) = B(t - \tau)\sigma(\tau)$.

Summarizing the above, we conclude that on the whole time axis, there holds

$$u(t,\tau) = B(t-\tau)\sigma(\tau). \tag{3.15}$$

From the first equation of (3.13), we obtain

$$S(t) = S(t_0)e^{-\int_{t_0}^t F(\xi)d\xi} = S(t_0)\mathcal{A}(t;t_0) \tag{3.16}$$

with

$$\mathcal{A}(t;t_0) := e^{-\int_{t_0}^t F(\xi)d\xi}$$

representing the probability for an uninfected individual at time t_0 to escape from becoming infected at least until time t. Furthermore, the cumulative force of infection (CFOI) during the period $[t_0, t]$ is given as

$$Y(t; t_0) := \int_{t_0}^{t} F(\xi) d\xi. \tag{3.17}$$

According to (3.16), given that $S(t_0) \approx N_0$ (the total population), we actually have the following relation among the CFI $Y(t; t_0)$, "escaping" probability $\mathcal{A}(t; t_0)$ and $s_\infty := s(\infty) = S(\infty)/N_0$:

$$s_\infty = \mathcal{A}(\infty; t_0) = e^{-Y(\infty; t_0)}. \tag{3.18}$$

Obviously, divergence of $Y(\infty; t_0)$ is equivalent to $s_\infty = 0$ from (3.18).

Denote

$$\hat{\mathcal{R}}_o(t) : = P(0) N_0 \int_{\tau}^{t} \varphi(a) da, \quad t \geq \tau. \tag{3.19}$$

Then the basic reproduction number of (3.13) is given by

$$\mathcal{R}_o := \hat{\mathcal{R}}_o(\infty).$$

The following theorem simply means that the disease will eventually die out. This is not surprising because the model does not have a demographic structure, and hence, there is no recruitment for the susceptible class.

Theorem 3.1 For model (3.14), the total infected population $U(\infty) = 0$ holds. Accordingly, $E(\infty) = 0, I(\infty) = 0$.

Proof We leave the proof in the Appendix. □

Because of the results in Theorem (3.1), it is reasonable to assume the following for the severity variable:

Assumption 3.1 $L(\infty) = 0$ and $L'(\infty) = 0$; and $U = 0 \leftrightharpoons L = 0$.

4 Final epidemic sizes for the model (3.14)

4.1 The case that the initial condition is given by $S(-\infty)$:

To make mathematical analysis easier, Breda et al. (2012), Diekmann et al. (2021) (references therein) assume time starts from $-\infty$. This assumption gives some convenience in the analysis due to the occurrence of the related integrals in the REs. In this subsection, we adopt this assumption on the starting time, i.e. $t_0 = -\infty$, so that

$B(t)$ and $F(t)$ have the explicit expressions (3.10) and (3.11), respectively. Further, we also accordingly impose the following initial conditions:

Assumption 4.1 $\lim_{t\to-\infty} S(t) = N_0$ and $\lim_{t\to-\infty} U(t) = 0$.

This assumption means the total infected population U was negligible in the infinite past. Under the Assumptions (3.1) and (4.1), there are hold

$$\lim_{t\to\pm\infty} P_l(t) = P(L(\pm\infty)) = P(0), \lim_{t\to\pm\infty} P_l'(t) = P'(L)L' = 0 \text{ and } \lim_{t\to\pm\infty} F(t) = 0. \tag{4.1}$$

It follows from the expression (3.10) that

$$\begin{aligned} F(t) &= P_l(t)\int_\tau^\infty \varphi(a)B(t-a)da \\ &= P_l(t)\int_\tau^\infty F(t-a)S(t-a)\varphi(a)da \\ &= P_l(t)\int_{-\infty}^{t-\tau} F(\eta)S(\eta)\varphi(t-\eta)d\eta. \end{aligned} \tag{4.2}$$

The cumulative force-of-infection over $(-\infty, T)$ is

$$\begin{aligned} Y(T;-\infty) &= \int_{-\infty}^T F(t)dt \\ &= -\int_\tau^{+\infty}\varphi(a)\left\{\int_{-\infty}^T \frac{dS(t-a)}{dt}P_l(t)dt\right\}da \\ &= -\int_\tau^{+\infty}\varphi(a)\left\{P_l(t)S(t-a)|_{t=-\infty}^{t=T} - \int_{-\infty}^T S(t-a)\frac{dP_l(t)}{dt}dt\right\}da \\ &= \hat{\mathcal{R}}_p(T) + S(-\infty)\int_\tau^{+\infty}\varphi(a)\left\{P_l(-\infty) - P_l(T)e^{-Y(T-a;-\infty)}\right\}da \\ &= \hat{\mathcal{R}}_p(T) + \mathcal{R}_o - P_l(T)\hat{\mathcal{H}}(T) \end{aligned} \tag{4.3}$$

where

$$\hat{\mathcal{R}}_p(t) := \int_{-\infty}^t \hat{\mathcal{H}}(\xi)P_l'(\xi)d\xi, \quad \mathcal{R}_p := \hat{\mathcal{R}}_p(\infty) \tag{4.4}$$

with

$$\hat{\mathcal{H}}(t) := \int_\tau^{+\infty}\varphi(a)S(t-a)da = \int_{-\infty}^{t-\tau}\varphi(t-\eta)S(\eta)d\eta. \tag{4.5}$$

Thus,

$$\hat{\mathcal{R}}'_P(t) = P'_l(t)\hat{\mathcal{H}}(t), \tag{4.6}$$

which indicates $P_l(t)$ and $\hat{\mathcal{R}}_p(t)$ have same direction of change.

Differentiating $\hat{\mathcal{H}}(t)$ gives

$$\hat{\mathcal{H}}'(t) = \int_\tau^\infty \frac{dS(t-a)}{dt}\varphi(a)da = -\int_\tau^\infty B(t-a)\varphi(a)da = -C(t) < 0; \tag{4.7}$$

We may also use the alternative formula in (4.5) to obtain

$$\hat{\mathcal{H}}'(t) = -\int_\tau^\infty \varphi(a)\frac{dS(t-a)}{da}da = S(t-\tau)\varphi(\tau) + \int_\tau^\infty \varphi'(a)S(t-a)da. \tag{4.8}$$

Here, we should mention that only for very special $\varphi(a)$ do the (4.6) and (4.8) reduce to explicit differential equations (DEs). For some details on this, see Sect. 5.

Also, a simple calculation leads to

$$\begin{aligned} &P_l(-\infty)\hat{\mathcal{H}}(-\infty) = P_l(-\infty)\int_\tau^{+\infty} \varphi(a)S(-\infty)da = \mathcal{R}_o; \\ &P_l(\infty)\hat{\mathcal{H}}(\infty) = \mathcal{R}_o s_\infty. \end{aligned} \tag{4.9}$$

Summarizing the above, the properties of $\hat{\mathcal{H}}(t)$ are given in the following

Lemma 4.1 $\hat{\mathcal{H}}(t)$ given in (4.5) is strictly decreasing on $t \in (-\infty, \infty)$ with $\hat{\mathcal{H}}(-\infty) = \frac{\mathcal{R}_o}{P(0)}$ and $\hat{\mathcal{H}}(\infty) = \frac{\mathcal{R}_o s_\infty}{P(0)}$.

Before discussing the property of $\mathcal{R}_p$, we give the following Lemma.

Lemma 4.2 Assume that f and g are functions defined on $\mathbb{R}$, satisfying

(i) $g : \mathbb{R} \to (g_m, g_M) \subseteq (0, \infty)$ be a differentiable, bounded and monotonic function with

$$\inf_{t\in\mathbb{R}} g(t) := g_m, \sup_{t\in\mathbb{R}} g(t) := g_M;$$

(ii) $f : \mathbb{R} \to (0, f_M)$ be a differentiable and bounded function satisfying

$$f(\pm\infty) = \sup_{t\in\mathbb{R}} f(t) := f_M < \infty.$$

Then

$$\int_{-\infty}^{\infty} g(t)f'(t)dt$$

converges. Moreover, if g is strictly decreasing, then

$$(g_m - g_M)f_M \leq \int_{-\infty}^{+\infty} g(t)f'(t)dt \leq 0. \tag{4.10}$$

Proof See Appendix. □

By the property of function $P_l(t)$ (i.e. $P(L)$), Lemmas 4.1 and 4.2, we immediately have the following theorem.

Theorem 4.1 For $\mathcal{R}_p$ is defined in (4.4), there holds

$$-\mathcal{R}_o < \mathcal{R}_p \leq 0. \tag{4.11}$$

By (4.3), the CFOI at $t = \infty$ with $t_0 = -\infty$ is

$$Y(\infty; -\infty) = \ln\left(\frac{1}{s_\infty}\right) = \mathcal{R}_p + \mathcal{R}_o\left(1 - e^{-Y(\infty;-\infty)}\right). \tag{4.12}$$

That is

$$-\ln(s_\infty) = \mathcal{R}_p + \mathcal{R}_o(1 - s_\infty); \tag{4.13}$$

equivalently,

$$r_\infty = \frac{Y(\infty; -\infty)}{\mathcal{R}_o} + \frac{\mathcal{R}_p}{\mathcal{R}_o}. \tag{4.14}$$

Remark 4.1 Compared with (1.8) and (1.9) which depend on a single parameter $\mathcal{R}_0$, (4.12) and (4.14) depend not only on parameter $\mathcal{R}_o$ but also $\mathcal{R}_p$. Note that $\mathcal{R}_p$ accounts for the impact of the fraction function $P_l(t) = P(L(t))$, which reflects the behaviour changes during an epidemic. To some sense, $\mathcal{R}_p$ captures the cumulative effect of behavioural adaptation/or NPIs over the entire infection window $(-\infty, \infty)$ on the final size, reflecting both its direction (i.e., whether the effect is beneficial or adverse, via the sign) and strength (via the magnitude); Theorem 4.1 shows that $\mathcal{R}_p < 0$ confirming the beneficial effect of behavioural adaption/or NPIs, as intuitively expected.

Rearranging the final size equation (4.13) yields

Theorem 4.2 Given that $S(-\infty) \approx N_0$, the final size relation (FSR) of model (3.14) is explicitly presented in the form

$$s_\infty = e^{-\mathcal{R}_o(1-s_\infty)-\mathcal{R}_p}; \tag{4.15}$$

or equivalently,

$$r_\infty = 1 - e^{-\mathcal{R}_p} e^{-\mathcal{R}_o r_\infty} \tag{4.16}$$

where $\hat{\mathcal{R}}_p(t)$ and $\mathcal{R}_p$ are defined in (4.4) and s_∞ and r_∞ are as in (2.7).
From the formula (4.15), we can see the fraction s_∞ is necessarily the intersection point of the following functions

$$q_1(x) = e^{-\mathcal{R}_o(1-x)}, \quad q_2(x) = e^{\mathcal{R}_p} x \tag{4.17}$$

in the open interval (0, 1). Note that $q_1(x)$ is an exponentially growing function and $q_2(x)$ is an linearly growing function with slope $e^{-\mathcal{R}_p}$ depending on $\mathcal{R}_p$.

Observe that $q_2(x)$ is increasing in the parameter $\mathcal{R}_p$, and $q_2(x) = x$ when $\mathcal{R}_p = 0$. This together with the fact that $\mathcal{R}_0 \leq 0$ implies the following

(C1) If $\mathcal{R}_o \leq 1$, $q_1(x) > x$ on $x \in [0, 1)$, that is $q_1(x) > q_2(x)$ for any $\mathcal{R}_p \leq 0$. It means there is only one $s_\infty = 1 \in [0, 1]$. Therefore, in this case, the final size r_∞ is close to zero. The same conclusion is also obtained in Brauer (2005), Breda et al. (2012) without $\mathcal{R}_p$ term.

(C2) If $\mathcal{R}_o > 1$, there exists a critical value $\mathcal{R}_p^{cr}$ for critical value $\mathcal{R}_p$ such that

(i) when $\mathcal{R}_p \in (\mathcal{R}_p^{cr}, 0)$, $q_1(x)$ and $q_2(x)$ have two intersect points $x_1^* \in (0, 1/\mathcal{R}_o)$ and $x_2^* \in (1/\mathcal{R}_o, 1)$;
(ii) when $\mathcal{R}_p = \mathcal{R}_p^{cr}$, there is only one intersect point $x^* = 1/\mathcal{R}_o \in (0, 1)$;
(iii) when $\mathcal{R}_p = 0$, there is only one intersect point $x^* \in (0, 1/\mathcal{R}_o)$.

The critical value $\mathcal{R}_p^{cr}$ is indeed determined by the two tangential conditions $q_1(x) = q_2(x)$ and $q_1'(x) = q_2'(x)$ which turns out to be

$$\mathcal{R}_p^{cr} := \ln(\mathcal{R}_o) - \mathcal{R}_o + 1$$

which satisfies

$$-\frac{(1-\mathcal{R}_o)^2}{\mathcal{R}_o} < \mathcal{R}_p^{cr} \leq 0$$

for any $\mathcal{R}_o > 1$. And

$$q_2^{cr}(x) := e^{\mathcal{R}_p^{cr}} x \tag{4.18}$$

is the tangent line of $q_1(x)$ with the tangent point $(1/\mathcal{R}_o, e^{1-\mathcal{R}_o})$. See Fig. 2 for a illustration.

Remark 4.2 $P_l(t) \equiv 1$ indicates $\mathcal{R}_p = 0$. In this special case, C2-(ii)–(iii) shows s_t will eventually fall down to a level s_∞ below $1/\mathcal{R}_o$, provided that $\mathcal{R}_o > 1$.

Remark 4.3 For the case (C2)-(i), there are two intersection values x_1^* and x_2^*. Remark 4.2 seems to suggest that x_2^* should be excluded and x_1^* is the true value of s_∞. Another intuition to support this suggestion is that x_2^* is not s_∞ is that $\mathcal{R}_p$ impact x_1^* and x_2^* in totally opposite direction: the larger $|\mathcal{R}_p|$ is (note $\mathcal{R}_p \le 0$), the larger x_1^* is and the smaller x_2^* is. Noting that $\mathcal{R}_p$ accounts for the effect of the intervention or human behaviour changes and precaution induce the fraction $P_l(t)$ should help mitigate an epidemic and hence increase the final since s_∞. Unfortunately, we cannot exclude x_2^* at present and can only leave it as a conjecture, as stated below.

Conjecture 4.1 If $\mathcal{R}_o > 1$, the final size s_∞ satisfies $s_\infty \le \frac{1}{\mathcal{R}_o}$.

It is difficult to prove the conjecture 4.1 here for a general case which is plausible. Fortunately, we can prove this conjecture for some special cases, as is done in Theorem 5.1 in the next Section. Notably, in classical SIR models without behavioural change, the conclusion of Conjecture 4.1 is both mathematically provable and biologically sound. And $\frac{1}{\mathcal{R}_o}$ is the so-called *normalized herd immunity threshold*, i.e., the peak of new infections (Fine et al. 2011). The inequality $s_\infty \le \frac{1}{\mathcal{R}_o}$ reflects that active infections still generate secondary cases at a declining rate even after herd immunity is achieved until removal or recovery, driving the susceptible population below the threshold before the epidemic ends.

Assume that the above conjecture is true, then the dependence of $s_\infty = x_1^*$ on $\mathcal{R}_p$ leads to the following lemma.

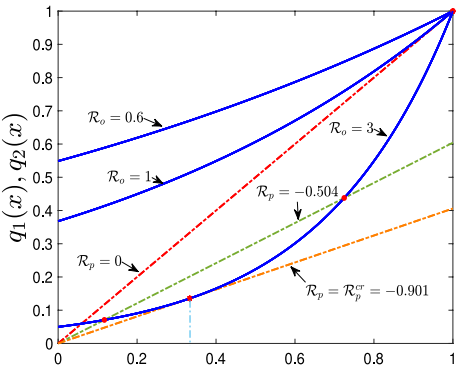

Fig. 2 Demonstration of $q_1(x)$ and $q_2(x)$

Theorem 4.3 (Effectiveness via $|\mathcal{R}_p|$) Suppose $\mathcal{R}_o > 1$ is fixed and $s_\infty \leq \frac{1}{\mathcal{R}_o}$, then the decrement of $\mathcal{R}_p$(i.e. increment of $|\mathcal{R}_p|$) leads to the increment of s_∞, i.e. the decrement of the final size r_∞. Indeed, the decrement of $\mathcal{R}_p$ from 0 to a negative value leads to the increment of s_∞ with the rate $1/\ |\mathcal{R}_o - \frac{1}{s_\infty}\ |$.

Proof See the Appendix. □

As described in Remark 4.1, $\mathcal{R}_p$ represents the cumulative effect of behaviour change on final size over the entire infection window, and $\mathcal{R}_p < 0$ always hold (see Theorem 4.1), indicating a beneficial impact on control. Theorem 4.3 shows that $|\mathcal{R}_p|$ measures intervention effectiveness: a greater $|\mathcal{R}_p|$ leads to a larger fraction of individuals escaping infection under certain conditions.

We have mentioned that adopting the initial time to be at $t_0 = -\infty$ brings in some convenience in analysis. However, it has an inconvenience for numerical computation. For convenience simulations that will be done in Sect.5, we explore the scenario of adopting initial time at $t_0 = 0$ in the next subsection.

4.2 The case with initial condition given by *S*(0)

Denote $L(0) := L_0$. According to the property of $P(L)$ described in the Assumption 1.3, we have

$$P_l(0) = P(L_0), \quad P_l(\infty) = P(L(\infty)) = P(0).$$

For $t \geq \tau$,

$$\begin{aligned} F_1(t) &= P_l(t)\left(\int_\tau^t \varphi(a)B(t-a)da + G(t)\right) \\ &= P_l(t)\left(\int_\tau^t -\frac{dS(t-a)}{dt}\varphi(a)da + G(t)\right) \\ &= P_l(t)\left(\int_\tau^t \frac{dS(t-a)}{da}\varphi(a)da + G(t)\right) \\ &= P_l(t)\left(S(0)\varphi(t) - S(t-\tau)\varphi(\tau) - \int_\tau^t S(t-a)\varphi'(a)da + G(t)\right) \end{aligned} \tag{4.19}$$

where

$$\frac{dS(t-a)}{dt} = -\frac{dS(t-a)}{da} = -F(t-a)S(t-a).$$

For $0 \leq t < \tau$

$$F_2(t) = P_l(t)\left(\int_\tau^\infty \beta(a)u(0,a-t)\frac{\sigma(a)}{\sigma(a-t)}da\right) = P_l(t)G_\tau(t). \tag{4.20}$$

In summary,

$$F(t) = \begin{cases} F_1(t) & t \geq \tau; \\ F_2(t) & 0 \leq t < \tau. \end{cases} \tag{4.21}$$

In addition

$$\lim_{t\to\tau} F_2(t) = P_l(\tau)\left(\int_\tau^\infty \beta(a)u(0,a-\tau)\frac{\sigma(a)}{\sigma(a-\tau)}da\right) = P_l(\tau)G(\tau) = F_1(\tau), \tag{4.22}$$

Thus, $F(t)$ is continuous.

Analogously to (4.5) and (4.4), define the following function

$$\hat{\mathcal{H}}_0(t) := \begin{cases} \int_\tau^t \varphi(a)S(t-a)da = \int_0^{t-\tau} \varphi(t-\eta)S(\eta)d\eta & t \geq \tau \\ 0 \quad t < \tau \end{cases} \tag{4.23}$$

$$\check{\mathcal{R}}_p(t) := \int_\tau^t \hat{\mathcal{H}}_0(a)P_l'(a)da, \quad t \geq \tau \tag{4.24}$$

$$\hat{\mathcal{R}}_c(t) := N_0 \int_\tau^t P_l(a)\varphi(a)da, \quad t \geq \tau \tag{4.25}$$

$$\hat{\mathcal{R}}_g(t) := \int_\tau^t P_l(a)G(a)da, \quad t \geq \tau \tag{4.26}$$

$$\hat{\mathcal{R}}_{op}(t) := \frac{S_0}{N_0}\hat{\mathcal{R}}_c(t) + \check{\mathcal{R}}_p(t) + \hat{\mathcal{R}}_g(t), \quad t \geq \tau \tag{4.27}$$

and use the notations

$$\mathcal{H}_0 := \hat{\mathcal{H}}_0(\infty),\ \mathcal{R}_{pp} := \check{\mathcal{R}}_p(\infty),\ \mathcal{R}_c := \hat{\mathcal{R}}_c(\infty),\ \mathcal{R}_g := \hat{\mathcal{R}}_g(\infty),\ \mathcal{R}_{op} := \hat{\mathcal{R}}_{op}(\infty). \tag{4.28}$$

The following properties hold:

$$\hat{\mathcal{H}}_0(\tau) = 0, \quad \mathcal{H}_0 = \frac{\mathcal{R}_o s_\infty}{P(0)}, \quad P_l(\infty)\mathcal{H}_0 = \mathcal{R}_o s_\infty,$$

$$\hat{\mathcal{R}}_c(\tau) = 0,\ \check{\mathcal{R}}_p(\tau) = 0,\ \hat{\mathcal{R}}_g(\tau) = 0,\ \hat{\mathcal{R}}_{op}(\tau) = 0,$$

and

$$\frac{\mathcal{R}_c}{\mathcal{R}_o} = \frac{\int_\tau^\infty P_l(a)\varphi(a)da}{P(0)\int_\tau^\infty \varphi(a)da} \leq 1.$$

Differentiating $\mathcal{H}_0(t)$ leads to

$$\begin{aligned}\mathcal{H}_0'(t) = \varphi(t)S(0) + \int_\tau^t \varphi(a)\frac{dS(t-a)}{dt}da = \varphi(t)S(0) - \left(\frac{F_1(t)}{P(t)} - G(t)\right)\\ = \varphi(\tau)S(t-\tau) + \int_\tau^t \varphi'(a)S(t-a)da.\end{aligned} \tag{4.29}$$

The sign of $\mathcal{H}_0'(t)$ is not clear. Unfortunately, it seems not direct to get the boundary of $\mathcal{R}_{pp}$, which is different from the $\mathcal{R}_p$ defined by (4.4) in Sect. 4.1.

Assume that $T > \tau > 0$,

$$\begin{aligned}Y(T;0) := \ln\left(\frac{S(0)}{S(T)}\right) = \int_0^T F(t)dt = \left(\int_0^\tau F_2(t)dt + \int_\tau^T F_1(t)dt\right)\\ = Y(\tau;0) + Y(T;\tau) = Y_0(T;0) + \frac{S_0}{N_0}\hat{\mathcal{R}}_c(T) - P_l(T)\hat{\mathcal{H}}_0(T) + \check{\mathcal{R}}_p(t)\end{aligned} \tag{4.30}$$

where

$$\begin{aligned}Y_0(T;0) &:= \int_0^T P_l(t)G^\diamond(t)dt = \int_0^\tau P_l(t)G_\tau(t)dt + \int_\tau^T P_l(t)G(t)dt\\ &= Y(\tau;0) + \hat{\mathcal{R}}_g(t)\\ &= \int_\tau^T \varphi(a)\left\{\int_0^a \frac{P_l(t)u_0(a-t)}{\sigma(a-t)}dt\right\}da\\ &\quad + \int_T^\infty \varphi(a)\left\{\int_0^T \frac{P_l(t)u_0(a-t)}{\sigma(a-t)}dt\right\}da\\ &= \int_\tau^\infty \varphi(a)\left\{\int_0^T \frac{P_l(t)u_0(a-t)}{\sigma(a-t)}dt\right\}da\\ &\quad - \int_\tau^T \varphi(a)\left\{\int_a^T \frac{P_l(t)u_0(a-t)}{\sigma(a-t)}dt\right\}da.\end{aligned} \tag{4.31}$$

with

$$Y(\tau;0) = \int_0^\tau P_l(t)G_\tau(t)dt = \int_0^\tau P_l(t)\left\{\int_{\tau-t}^\infty u_0(\eta)\frac{\varphi(\eta+t)}{\sigma(\eta)}d\eta\right\}dt. \tag{4.32}$$

$Y_0(T;0)$ is the something that occurred before time $t = 0$.

Letting T tend to infinity, we get

$$\begin{aligned} Y_0(\infty;0) &= \int_\tau^\infty \varphi(a)\left\{\int_0^\infty \frac{P_l(t)u_0(a-t)}{\sigma(a-t)}dt\right\}da \\ &\quad - \int_\tau^\infty \varphi(a)\left\{\int_a^\infty \frac{P_l(t)u_0(a-t)}{\sigma(a-t)}dt\right\}da \\ &= \int_\tau^\infty \varphi(a)\left\{\int_0^a \frac{P_l(a-\eta)u_0(\eta)}{\sigma(\eta)}d\eta\right\}da. \end{aligned} \tag{4.33}$$

Furthermore,

$$Y(t;\tau) = \hat{\mathcal{R}}_g(t) + \frac{S_0}{N_0}\hat{\mathcal{R}}_c(t) - P_l(t)\hat{\mathcal{H}}_0(t) + \check{\mathcal{R}}_p(t) = \hat{\mathcal{R}}_{op}(t) - P_l(t)\hat{\mathcal{H}}_0(t) \tag{4.34}$$

and

$$\begin{aligned} Y(\infty;0) &= Y_0(\infty;0) + S(0)\int_\tau^\infty \varphi(a)P_l(a)\left\{1 - \frac{P_l(\infty)}{P_l(a)}e^{-Y(\infty;0)}\right\}da \\ &\quad + \int_\tau^\infty \left\{\int_\tau^t \varphi(a)S(t-a)da\right\}P_l'(t)dt \\ &= Y_0(\infty;0) + \mathcal{R}_{pp} + \frac{S_0}{N_0}\mathcal{R}_c - P_l(\infty)\mathcal{H}_0 \\ &= \mathcal{R}_{pp} + \left(1 - \frac{s_\infty}{N_0}\right)\mathcal{R}_o - \int_\tau^\infty \varphi(a)\left\{P(0)N_0 - S(0)P(I(a))\right. \\ &\quad \left. - \int_0^a \frac{P_l(t)u_0(a-t)}{\sigma(a-t)}dt\right\}da \\ &= \mathcal{R}_{pp} + \frac{S_0}{N_0}\left(\frac{\mathcal{R}_c}{\mathcal{R}_o} - e^{-Y(\infty;0)}\right)\mathcal{R}_o + Y_0(\infty;0). \end{aligned} \tag{4.35}$$

Correspondingly,

$$Y(\infty;\tau) = \mathcal{R}_{op} - \mathcal{R}_o\frac{S(\tau)}{N_0}e^{-Y(\infty;\tau)}. \tag{4.36}$$

We collect our findings into the following theorem.

Theorem 4.4 Given the initial condition $S(0) = S_0$, the final size relation of the model (3.14) is

$$Y(\infty;0) = \mathcal{R}_{pp} + \frac{S_0}{N_0}\left(\frac{\mathcal{R}_c}{\mathcal{R}_o} - e^{-Y(\infty;0)}\right)\mathcal{R}_o + Y_0(\infty;0) \tag{4.37}$$

where $Y_0(\infty;0)$ is given in (4.33). Moreover, the following form also holds:

$$Y(\infty;\tau) = \mathcal{R}_{op} - \mathcal{R}_o \frac{S(\tau)}{N_0} e^{-Y(\infty;\tau)};$$

equivalently,

$$0 = \ln(s_\infty) + (\mathcal{R}_{op} - \mathcal{R}_o) + \mathcal{R}_o r_\infty - \ln(s(\tau)).$$

where $\hat{\mathcal{R}}_{op}(t)$ and $\mathcal{R}_{op}$ are given in (4.27) and (4.28).

Remark 4.4 In Brauer (2008) and Breda et al. (2012), $Y_0(\infty;0)$ is set to 0. Letting $Y_0(\infty;0) \downarrow 0$ in (4.35), we obtain

$$Y(\infty;0) = \mathcal{R}_{pp} + \frac{S_0}{N_0}\left(\frac{\mathcal{R}_c}{\mathcal{R}_o} - e^{-Y(\infty;0)}\right)\mathcal{R}_o \le \mathcal{R}_{pp} + \frac{S_0}{N_0}\left(1 - e^{-Y(\infty;0)}\right)\mathcal{R}_o, \quad (4.38)$$

Note that $\mathcal{R}_g$ reflects the contribution from those already in the infected class at time $t = 0$ to the cumulative force-of-infection through the epidemic. When the initial fraction of infectors $\frac{N_0 - S_0}{N_0}$ is small, then $\mathcal{R}_g$ is small and hence, can be safely omitted. Thus,

$$\hat{\mathcal{R}}_{op}(t) - \check{\mathcal{R}}_p(t) \approx \hat{\mathcal{R}}_c(t) \quad (4.39)$$

which leads to the following result:

Corollary 4.1 If $\frac{N_0 - S_0}{N_0}$ is small i.e. $\frac{N_0 - S_0}{N_0} = \epsilon$ for $0 < \epsilon \ll 1$, then

$$Y(\infty;0) = \mathcal{R}_{pp} + \left(\frac{\mathcal{R}_c}{\mathcal{R}_o} - e^{-Y(\infty;0)}\right)\mathcal{R}_o$$

or

$$-\ln(s_\infty) \approx \mathcal{R}_{pp} + \mathcal{R}_o\left[\frac{\mathcal{R}_c}{\mathcal{R}_o} - s_\infty\right] < \mathcal{R}_{pp} + \mathcal{R}_o r_\infty.$$

In what follows, we consider a special case of “survival" probability (probability of staying in the infected class). For such a special case, we can obtain more explicit and possibly more useful results.

Consider

$$\gamma(a) = \begin{cases} \gamma_1, & a \in [0,\tau), \\ \gamma_2, & a \in [\tau,\infty). \end{cases} \quad (4.40)$$

With this piecewise constant recovery rate function $\gamma(a)$, the “survival" probability function $\sigma(a)$ becomes

$$\sigma(a) = \begin{cases} e^{-\gamma_1 a}, & a \in [0, \tau) \\ V_\tau e^{-\gamma_2 a}, & a \in [\tau, \infty) \end{cases} \tag{4.41}$$

where

$$V_\tau := V(\tau) = e^{(\gamma_2 - \gamma_1)\tau}$$

with

$$V(t) = e^{(\gamma_2 - \gamma_1)t}, \quad t \in [0, \tau]. \tag{4.42}$$

Accordingly,

$$T_E = \frac{1}{\gamma_1} - \frac{e^{-\gamma_1 \tau}\tau}{1 - e^{-\gamma_1 \tau}} \geq 0, \quad T_I = \frac{V_\tau(\gamma_2 \tau + 1)e^{-\gamma_2 \tau}}{\gamma_2 e^{-\gamma_1 \tau}} = \tau + \frac{1}{\gamma_2}.$$

Obviously,

$$\lim_{\gamma_1 \to 0} T_E = \frac{\tau}{2}, \quad \lim_{\gamma_1 \to \infty} T_E = 0.$$

Accordingly,

$$\varphi(a) = \begin{cases} 0, & a \in [0, \tau), \\ \beta(a) V_\tau e^{-\gamma_2 a}, & a \in [\tau, \infty). \end{cases} \tag{4.43}$$

By computing, we get

$$\begin{array}{ll} G_\tau(t) = \sigma(t) \int_{\tau - t}^{\tau} u_0(\eta)\beta(\eta + t)\frac{1}{V(\eta)} d\eta + \frac{\sigma(t)}{V_\tau} \int_\tau^\infty u_0(\eta)\beta(\eta + t) d\eta & t < \tau, \\ G(t) = \sigma(t) \int_0^\tau u_0(\eta)\beta(\eta + t)\frac{1}{V(\eta)} d\eta + \frac{\sigma(t)}{V_\tau} \int_\tau^\infty u_0(\eta)\beta(\eta + t) d\eta & t \geq \tau. \end{array} \tag{4.44}$$

In reality, the initial age density function $u_0(a)$ can have various situations. In the sequel, we will discuss two situations: continuous and discretely centred.

Assumption 4.2 The initial infection-age distribution $u_0(a)$ is continuous and satisfies

$$\int_0^\tau u_0(a) da = E_0, \quad \int_\tau^\infty u_0(a) da = I_0. \tag{4.45}$$

For the continuous initial density function, we have

Lemma 4.3 If $u_0(a)$ satisfies the Assumption 4.2, and

$$\varphi(a) = \begin{cases} 0, & a \in [0, \tau), \\ \beta V_\tau e^{-\gamma_2 a}, & a \in [\tau, \infty), \end{cases} \tag{4.46}$$

then the following final size inequality holds:

$$Y(\infty;\tau) \le \mathcal{R}_o \left(\max\left\{ \frac{1}{V_\tau}, 1 \right\} - \frac{S(\tau)}{N_0} e^{-Y(\infty;\tau)} \right) + \mathcal{R}_{pp}. \tag{4.47}$$

Especially, if $\gamma_1 = 0$,

$$Y(\infty;\tau) \le \mathcal{R}_o \left(1 - \frac{S(\tau)}{N_0} e^{-Y(\infty;\tau)} \right) + \mathcal{R}_{pp}. \tag{4.48}$$

Proof See the Appendix. □

Assumption 4.3 The initial infection-age distribution $u_0(a)$ is discretely centred, meaning that $u_0(a) = 0$ for all a except for finite many values of a.

For example, if all initial infectives have infection-age τ at $t = 0$, and all the infected but not infective have infection-age $\bar{a}_l \in (0, \tau)$ at $t = 0$ then the initial infection-age distribution is

$$u_0(a) = \begin{cases} I_0, & \text{for } a = \tau, \\ E_0, & \text{for } a = \bar{a}_l, \\ 0, & \text{Otherwise.} \end{cases} \tag{4.49}$$

The Assumption 4.3 on $u_0(a)$ indicates $G^\diamond(t) \equiv 0$, $\hat{\mathcal{R}}_g(t) \equiv 0$ and $Y_0(t;0) \equiv 0$, $Y(\tau;0) \equiv 0$. Further,

$$\hat{\mathcal{R}}_{op}(t) = \frac{S_0}{N_0} \hat{\mathcal{R}}_c(t) + \check{\mathcal{R}}_p(t), \tag{4.50}$$

Therefore, we have

$$Y(t;0) = Y(t;\tau) = \hat{\mathcal{R}}_{op}(t) - P_l(t)\hat{\mathcal{H}}_0(t), \tag{4.51}$$

and thus, the limit of $Y(t,\tau)$ as $t \to \infty$ reduces to

$$\begin{aligned} Y(\infty;\tau) &= \mathcal{R}_{pp} + \left(1 - \frac{S(\tau)}{N_0} e^{-Y(\infty;\tau)} \right) \mathcal{R}_o \\ &\quad - \int_\tau^\infty \varphi(a) \left\{ P(0)N_0 - S(0)P(I(a)) \right\} da \\ &= \mathcal{R}_{pp} + \left(\frac{S_0}{N_0} \frac{\mathcal{R}_c}{\mathcal{R}_o} - \frac{S(\tau)}{N_0} e^{-Y(\infty;\tau)} \right) \mathcal{R}_o \le \mathcal{R}_{pp} + \mathcal{R}_o \left(s_0 - s_\infty \right). \end{aligned} \tag{4.52}$$

Summarizing the above, we have the following conclusion.

Lemma 4.4 If $u_0(a)$ satisfies the Assumption 4.3, then the final size relation is,

$$-\ln(s_\infty) = \mathcal{R}_{pp} + \frac{S_0}{N_0}\mathcal{R}_c - \mathcal{R}_o s_\infty - \ln\left(\frac{S(\tau)}{N_0}\right). \tag{4.53}$$

Remark 4.5 If $P_l(t) \equiv 1$, then

$$\check{\mathcal{R}}_p(t) \equiv 0,\ \hat{\mathcal{R}}_c(t) \equiv \hat{\mathcal{R}}_o(t).$$

Moreover,

$$\mathcal{R}_c = \mathcal{R}_o,\ \mathcal{R}_{op} = \frac{S_0}{N_0}\mathcal{R}_o,$$

and the final size equation (4.53) for s_∞ is transformed to the formula (1.1) for r_∞. We point out that when $P_l(t) \le 1$, in our numerical simulations for the special models presented in the next section, we observe that the inequality $\mathcal{R}_{op} \le \mathcal{R}_o$ (corresponding to Theorem 4.1 in Sect. 4.1) always holds. See, Figs. 3b and 7. Unfortunately, we are unable to prove this inequality for the general model.

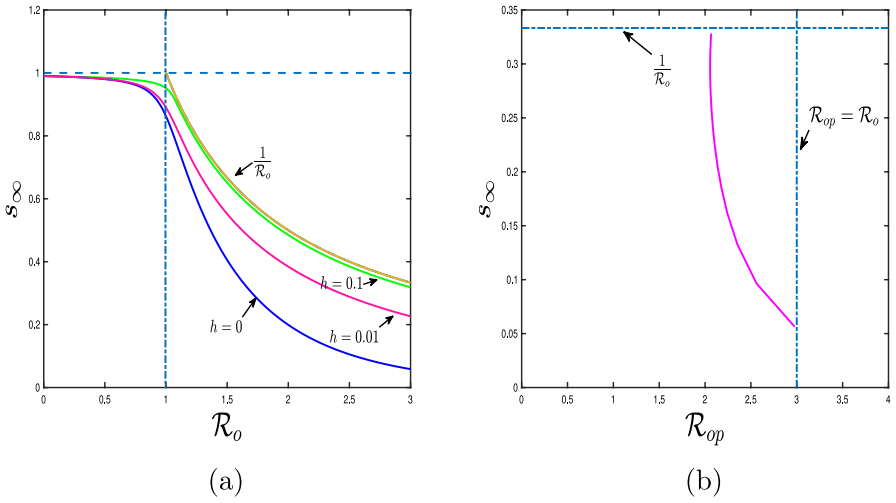

Fig. 3 **a** $\mathcal{R}_o \in [0, 3]$; **b** $h \in [0, 1], \mathcal{R}_o = 3$. Other parameters are $\tau = 4, \gamma_1 = 0, \gamma_2 = \frac{1}{7}, N_0 = 300, E_0 = 5, I_0 = 5, R(0) = 0$

5 Some applications and numerical simulations: reduction to DEs

In this section, we demonstrate the theoretical results established in the preceding sections by some particular simple examples that can reduce the model to DE models. We will also present some numerical simulation results to illustrate the theoretical results.

We begin with, as a preparation, the following Lemma, which is parallel to the Lemma 4.2 but is only on the half line $[0, \infty)$.

Lemma 5.1 Assume that

(i) $g : [0, \infty) \to [0, \infty)$ is differentiable, bounded and strictly increasing with $g(0) = 0$ (hence $g(t) \in [0, g_M)$ where $g_M := \sup_{t\in[0,\infty)} g(t) = g(\infty)$);

(ii) $f : [0, \infty) \to [0, \infty)$ is differentiable and bounded, and satisfies $f(t) \le f(\infty) =: f_M < \infty$.

Then,

$$0 \le \int_0^\infty f'(t)g(t)dt \le g_M f_M. \tag{5.1}$$

Proof See the Appendix. □

We now choose two particular simple forms for the kernel function $\phi(a)$ to explore the impact of $P_l(t)$ on the final size, reflected by the relation between s_∞ and $\mathcal{R}_P$ or $\mathcal{R}_{op}$.

The first one is the one given in (4.43) with $\beta(a) = \beta$ (constant), which is rewritten below as

$$\varphi_1(a) := \begin{cases} 0, & a \in [0, \tau) \\ \beta e^{(\gamma_2-\gamma_1)\tau} e^{-\gamma_2 a}, & a \in [\tau, \infty), \end{cases} \tag{5.2}$$

For this kernel, The corresponding basic reproductive number is

$$\mathcal{R}_o = P(0)N_0 \int_\tau^\infty \varphi_1(a)da = \frac{\varphi_1(\tau)P(0)N_0}{\gamma_2} = \{P_L(0)N_0\beta\sigma(\tau)\}(T_I - \tau). \tag{5.3}$$

Remark 5.1 The special case of $\gamma_1 \to 0$ corresponds to the scenario that a pre-infectious individual cannot recover before getting infectious. In this case, the latent period τ makes no difference in $\mathcal{R}_o$.

The second form is taken from Champredon et al. (2018) and is given by

$$\varphi_2(a) := \begin{cases} \beta\frac{\epsilon}{\epsilon-\bar{\gamma}}(e^{-\bar{\gamma}a} - e^{-\epsilon a}), & \text{if} \quad \bar{\gamma} \neq \epsilon, \\ \beta\bar{\gamma}ae^{-\bar{\gamma}a}, & \text{if} \quad \bar{\gamma} = \epsilon, \end{cases} \qquad a \in (0,\infty). \tag{5.4}$$

With this kernel, the basic reproduction number of the model is

$$\mathcal{R}_o = \frac{\beta P(0) N_0}{\bar{\gamma}}. \tag{5.5}$$

5.1 The case $\varphi(a) = \varphi_1(a)$

For this kernel, $u(t,\tau)$ is determined to be

$$u(t,\tau) = B(t-\tau)\sigma(\tau) = \varphi(\tau)S_P(t-\tau)I(t-\tau). \tag{5.6}$$

The formulas (4.7) and (4.8) lead to

$$-\beta I(t) = \varphi(\tau)S(t-\tau) - \gamma_2\hat{\mathcal{H}}(t) \tag{5.7}$$

which establishes an explicit expression for $\hat{\mathcal{H}}(t)$ as

$$\hat{\mathcal{H}}(t) = \frac{\varphi(\tau)}{\gamma_2}S(t-\tau) + \frac{\beta}{\gamma_2}I(t) \tag{5.8}$$

Also, the model (3.14) for this kernel reduces to

$$\begin{cases} S'(t) = -\beta I(t)S_P(t) \\ E'(t) = \beta I(t)S_P(t) - \gamma_1 E - \varphi(\tau)I(t-\tau)S_P(t-\tau) \\ I'(t) = -\gamma_2 I + \varphi(\tau)I(t-\tau)S_P(t-\tau) \\ R'(t) = \gamma_1 E + \gamma_2 I \end{cases} \tag{5.9}$$

In summary, the integrated model reads

$$\begin{cases} S'(t) = -\beta I P_l(t) S \\ E'(t) = \beta I(t)S_P(t) - \gamma_1 E - \varphi(\tau)I(t-\tau)S_P(t-\tau) \\ I'(t) = -\gamma_2 I + \beta e^{-\gamma_1\tau}I(t-\tau)P(I(t-\tau))S(t-\tau) \\ \hat{\mathcal{R}}_P'(t) = P_l'(t)\hat{\mathcal{H}}(t). \end{cases} \tag{5.10}$$

For the scenario of $t_0 = -\infty$, the initial condition is

$$(S(-\infty), E(-\infty), I(-\infty), \hat{\mathcal{H}}(-\infty), \hat{\mathcal{R}}_p(-\infty)) = (N_0, 0, 0, \mathcal{R}_o/P(0), 0).$$

and for the scenario $t_0 = 0$, the initial condition is

$$S(0) = S_0, \quad E(0) = E_0 = \int_{-\infty}^{0} E'(t)dt, \quad I(0) = I_0 = \int_{-\infty}^{0} I'(t)dt \tag{5.11}$$

and

$$\hat{\mathcal{R}}_p(0) = \int_{-\infty}^{0} \hat{\mathcal{H}}(t) P_l'(t)dt \tag{5.12}$$

which is difficult to determine.

Although the theoretical results for the scenario that t starts from the infinite past described in Sect. 4.1 are more brief and straightforward, it is not easy to verify these conclusions by numerical simulation.

In the rest of this section, we will choose $P(L) = e^{-hL}$ with $h \geq 0$ and use the results in Sect. 4.2 to show the numerical results.

5.1.1 The case $\tau > 0$ with $t_0 = 0$

We consider a particular initial density specified below.

$$u_0(a) = \begin{cases} \dfrac{E_0}{\tau} \quad a \in [0, \tau), \quad a \in [0, \tau), \\ \text{integral on } [\tau, \infty) \text{ with contraint } \displaystyle\int_{\tau}^{\infty} u_0(\eta)d\eta = I_0. \end{cases} \tag{5.13}$$

Then for $t \in [0, \tau)$,

$$u(t, \tau) = \frac{E_0}{\tau}\sigma(t) = \frac{E_0}{\tau}e^{-\gamma_1 t}, \tag{5.14}$$

and, accordingly

$$G_\tau(t) = \begin{cases} \frac{\beta\sigma(t)}{V_\tau}\left(\frac{1-V(t)}{\gamma_1-\gamma_2}\frac{E_0}{\tau} + I_0\right) & \gamma_1 \neq \gamma_2, \\ \beta\sigma(t)\left(\frac{E_0 t}{\tau} + I_0\right) & \gamma_1 = \gamma_2. \end{cases} \tag{5.15}$$

$$G(t) = \begin{cases} \frac{\beta\sigma(t)}{V_\tau}\left(\frac{1-V_\tau}{\gamma_1-\gamma_2}\frac{E_0}{\tau} + I_0\right) & \gamma_1 \neq \gamma_2, \\ \beta\sigma(t)(E_0 + I_0) & \gamma_1 = \gamma_2. \end{cases} \tag{5.16}$$

$$\hat{\mathcal{R}}_g(t) = \begin{cases} \frac{1}{N_0 V_\tau}\left(\frac{1-V_\tau}{\gamma_1-\gamma_2}\frac{E_0}{\tau} + I_0\right)\hat{\mathcal{R}}_c(t) & \gamma_1 \neq \gamma_2, \\ \frac{E_0+I_0}{N}\hat{\mathcal{R}}_c(t) & \gamma_1 = \gamma_2. \end{cases} \tag{5.17}$$

The disease dynamics for $t \in [0, \tau]$ is then governed by the ODE system

$$\begin{cases} S'(t) = -\beta I(t) S_P(t) \\ E'(t) = \beta I(t) S_P(t) - \gamma_1 E - \frac{E_0}{\tau} e^{-\gamma_1 t} \\ I'(t) = -\gamma_2 I + \frac{E_0}{\tau} e^{-\gamma_1 t} \end{cases} \tag{5.18}$$

For $t \geq \tau$,

$$u(t,\tau) = \sigma(\tau) B(t-\tau) = \beta e^{-\gamma_1 \tau} I(t-\tau) S_P(t-\tau).$$

Plugging the above calculations into (5.9), we obtain

$$\begin{cases} S'(t) = -\beta I(t) S_P(t) \\ E'(t) = \beta I(t) S_P(t) - \gamma_1 E - \beta e^{-\gamma_1 \tau} I(t-\tau) S_P(t-\tau) \\ I'(t) = -\gamma_2 I + \beta e^{-\gamma_1 \tau} I(t-\tau) S_P(t-\tau) \\ \hat{\mathcal{H}}_0'(t) = \varphi(\tau) S(t-\tau) - \gamma_2 \hat{\mathcal{H}}_0(t) \\ \mathcal{R}_{op}'(t) = \left(\hat{\mathcal{H}}_0'(t) + \hat{\mathcal{H}}_0(t) \frac{P_l'(t)}{P_l(t)} + \beta I(t) \right) P_l(t) \end{cases} \quad t \in [\tau, \infty). \tag{5.19}$$

with the new and translated initial condition $(S(\tau), E(\tau), I(\tau), \hat{\mathcal{H}}_0(\tau), \hat{\mathcal{R}}_{op}(\tau)) = (S_\tau, E_\tau, I_\tau, 0, 0)$ with $(S(\tau), E(\tau), I(\tau)) = (S_\tau, E_\tau, I_\tau)$ determined by solving the ODE system (5.18) on $[0, \tau]$.

By simulating the system (5.19) with the particular $L(t) = I(t)$, it is easy to numerically verify the FSR given in the Theorem 4.4. Figure 3 shows the numeric results with the parameter values specified in the caption. From Fig. 3, we observe the following.

(a). If $\mathcal{R}_o$ is fixed, an increase in h from 0 increases s_∞ (i.e. decreases the final size r_∞); for fixed $h \geq 0$, s_∞ decrease as $\mathcal{R}_o$ increase; Further, $s_\infty < 1/\mathcal{R}_o$ holds, which confirms the Conjecture 4.1.
(b). $\mathcal{R}_o - \mathcal{R}_{op} > 0$. This inequality can be theoretically proved for the SIR model (5.20) (the case of $\tau = 0$), see Lemma 5.2 below. In addition, s_∞ decreases (or the final size r_∞ increases) as $\mathcal{R}_{op}$ increases.

Unfortunately, all the rigid proof of these aforementioned numerical results (shown in Fig. 3) is not easy for the model (5.19).

5.1.2 The special case $\tau \equiv 0$ and $t_0 = 0$

If $\tau \equiv 0$, the model (3.14) reduces to the SIR model

$$\begin{cases} S'(t) = -\beta I(t) S_P(t) \\ I'(t) = -\bar{\gamma} I + \beta I(t) S_P(t) \\ R'(t) = \bar{\gamma} I \end{cases} \tag{5.20}$$

for which the basic reproduction number is

$$\mathcal{R}_o = \frac{\beta P(0) N_0}{\bar{\gamma}}. \tag{5.21}$$

By simple calculations, we can obtain the explicit formulas for functions $\hat{\mathcal{H}}_0(t)$, $\hat{\mathcal{R}}_c(t)$ and $\check{\mathcal{R}}_p(t)$ as

$$\begin{aligned} \hat{\mathcal{H}}_0(t) &= \frac{\beta}{\gamma_2} S(t) + \frac{\beta}{\gamma_2} I(t) - \frac{\varphi(t) N_0}{\gamma_2} \\ &= \frac{\mathcal{R}_o}{P(0)}(1 - \sigma(t)) - \frac{\mathcal{R}_o}{P(0)} r(t) \end{aligned} \tag{5.22}$$

with $\hat{\mathcal{H}}_0(0) = 0$ and $\hat{\mathcal{H}}_0(\infty) = \mathcal{R}_o s_\infty / P(0)$;

$$\begin{aligned} \hat{\mathcal{R}}_c(t) &= \beta N_0 \int_0^t \sigma(a) P_l(a) da \\ &= -\frac{\mathcal{R}_o}{P(0)} \int_0^t \sigma'(a) P_l(a) da; \end{aligned} \tag{5.23}$$

and

$$\check{\mathcal{R}}_p(t) = -\hat{\mathcal{R}}_c(t) + \frac{\mathcal{R}_o}{P(0)} P_l(t)(1 - \sigma(t)) - \frac{\mathcal{R}_o}{P(0)} \int_0^t r(t) P_l'(t) dt. \tag{5.24}$$

Corresponding to the Assumption 4.2, we give the following assumption:

Assumption 5.1 $\int_0^\infty u(0, \eta) d\eta = I_0$.

Under Assumption 5.1, we obtain

$$G^\diamond(t) = G(t) = \beta \int_t^\infty u(0, a - t) e^{-\gamma_2 t} da = \beta e^{-\gamma_2 t} \int_0^\infty u_0(\eta) d\eta = \beta e^{-\gamma_2 t} I_0 = \varphi(t) I_0, \tag{5.25}$$

where $G(t)$ is the total infectivity of the infected population at $t = 0$ at time t.

$$Y_0(t; 0) = \hat{\mathcal{R}}_g(t) = \frac{I_0}{N_0} \hat{\mathcal{R}}_c(t) \tag{5.26}$$

Then

$$\begin{aligned} \hat{\mathcal{R}}_{op}(t) &= \hat{\mathcal{R}}_c(t) + \check{\mathcal{R}}_p(t) \\ &= \frac{\mathcal{R}_o}{P(0)} P_l(t)(1 - \sigma(t)) - \frac{\mathcal{R}_o}{P(0)} \int_0^t r(t) P_l'(t) dt \end{aligned} \tag{5.27}$$

Using the Lemma 5.1, we get

$$\int_0^t r(t)P_l'(t)dt > 0.$$

Then the following result holds

Lemma 5.2 For model (5.20) with $P_l(t) = P(I(t))$ under the Assumption 5.1, $\mathcal{R}_o - \mathcal{R}_{op} > 0$ holds, see Fig. 7.

According to the theorem (4.4), we get the following inequality of the FSR of the model (5.20)

$$-\ln(s_\infty) = \mathcal{R}_o r(\infty) + (\mathcal{R}_{op} - \mathcal{R}_o) - \ln(s_0) < \mathcal{R}_o r(\infty) - \ln(s_0), \tag{5.28}$$

which includes

$$\frac{Y(\infty, 0)}{\mathcal{R}_o} < r_\infty. \tag{5.29}$$

Based on numerical simulation, system (5.20) in this Subsect. 5.1.2 exhibits similar qualitative features to system (5.20) in Subsect. 5.2. For brevity, the related figures and results are omitted here and included in the Appendix. In fact, the model proposed in Sereno et al. (2022) is similar to (5.20). To some extent, the work (Sereno et al. 2022) validates our modelling and results through real-world applications.

5.2 The case $\varphi(a) = \varphi_2(a)$

If choosing the kernel $\varphi(a) = \varphi_2(a)$ and letting $\tau = 0$, the model (3.13) reduces to (Details in the Appendix)

$$\begin{cases} S'(t) = -\beta S_P I \\ E'(t) = -\epsilon E + \beta S_P I \\ I'(t) = -\bar{\gamma} I + \epsilon E \\ R'(t) = \bar{\gamma} I \end{cases} \tag{5.30}$$

For this kernel, the basic reproduction number $\mathcal{R}_o$ of (3.14) has the same expression as (5.3) with $\tau = 0$.

Functions $\hat{\mathcal{H}}(t)$ and $\hat{\mathcal{R}}_p(t)$ have the following forms:

$$\hat{\mathcal{H}}(t) = -\frac{\beta}{\bar{\gamma}} R(t), \quad \hat{\mathcal{R}}_p(t) = \int_0^t P_l'(a)\hat{\mathcal{H}}(a)da \tag{5.31}$$

Since

$$S_P(t) = P_l(t)S \le S(t) \le S(0), \ \ t \ge 0,$$

and accordingly,

$$s_P(t) := \frac{S_P(t)}{N_0} \leq s(t) \leq s(0), \quad t \geq 0.$$

Remark 5.2 It needs to be emphasized that we choose $L(t) = U(t)$ in this Subsect. 5.2 to make the following theoretical analysis much easier. However, in practice, U_0 is not easy to assess due to the difficulty of obtaining and tracking E_0 in the early stages of epidemics.

Calculations lead to

$$U'(t) = P_l(t)S\beta I - \bar{\gamma}I = \mathcal{R}_o\left(\frac{s_P(t)}{P(0)} - \frac{1}{\mathcal{R}_o}\right)\bar{\gamma}I \leq \mathcal{R}_o\left(\frac{s_P(t)}{P(0)} - \frac{1}{\mathcal{R}_o}\right)\bar{\gamma}U \quad (5.32)$$

$$U''(t) = \mathcal{R}_o\left(\frac{s_P'(t)}{P(0)}\right)\bar{\gamma}I + \mathcal{R}_o\left(\frac{s_P(t)}{P(0)} - \frac{1}{\mathcal{R}_o}\right)\bar{\gamma}I'. \quad (5.33)$$

With the above preparation for this kernel, we then obtain the following results for the reduced model (5.30).

Theorem 5.1 For system (5.30), s_∞ satisfies $s_\infty \leq \frac{1}{\mathcal{R}_o}$. Moreover, if $P_l(t) = P(U)$, $s_\infty < \frac{1}{\mathcal{R}_o}$.

Proof See the Appendix. □

Moreover,

Lemma 5.3 If $U'(t_m) = 0$, then $U''(t_m) \leq 0$.

Proof See the Appendix. □

Remark 5.3 Lemma 5.3 indicates that if U has a critical point, it must be a local maximum. Further, the continuous differentiability of U implies that $U(t)$ has, at most, a single peak $U(t_m)$. In other words, if $U'(0) < 0$ then $U_{max} = U(0)$; if $U'(0) > 0$, then there exits $t_m > 0$ such that $U'(t_m) = 0$ and $U_{max} = U(t_m)$ and t_m is the precisely the time at which the fraction of infected individuals reaches the peak $U_{\max}$. Unfortunately, we are unable to determine the peak size analytically.

In summary, we have the following Theorem.

Theorem 5.2 (Peak time and size) Consider $P_l(t) = P(U)$. Then, $U'(0) \geq 0$ (which is equivalent to $P_l'(0) \geq 0$) if and only if $s_P(0) \geq \frac{P(0)}{\mathcal{R}_o}$. Moreover,

(I) When $\mathcal{R}_o < 1$, then $U'(t) < 0$ for all $t > 0$, the maximum value of $U(t)$ is attained at $t_0 = 0$, that is, $U_{max} = I(0) + E(0)$;
(II) When $\mathcal{R}_o > 1$, then

(a) if $s_P(0) > \frac{P(0)}{\mathcal{R}_o}$, there exists a $t_m > 0$ such that $U'(t_m) = 0$. Accordingly, $s(t_m) = \frac{1}{\mathcal{R}_o}\frac{P(0)}{P(U_{max})}$ where $U_{max} = U(t_m)$;

(b) if $s_P(0) < \frac{P(0)}{\mathcal{R}_o}$, then $U'(t) < 0$ for all $t > 0$, the maximum value occurs at $t_0 = 0$, that is, $U_{max} = I_0 + E_0$.

Remark 5.4 (See Fig. 5) From Theorem (5.2)-(II), since t_m is the threshold for $U'(t)$, the value

$$S(t_m) = \frac{P(0)}{P(U_{max})}\frac{N_0}{\mathcal{R}_o}$$

is indeed the threshold for herd immunity (HIT), i.e. the value of S under which U can no longer increase (Fine et al. 2011). Thus, the involvement of $P_l(t)$ leads to the increment of herd immunity threshold compared with the classical herd immunity level $N_0/\mathcal{R}_o$, since $P(0)/P(U_{max}) > 1$.
The following result is a direct result of Lemma 5.1.

Theorem 5.3 (See Fig. 4b) For (5.30), $\mathcal{R}_p = \hat{\mathcal{R}}_p(\infty) < 0$ always holds.

Recall that the CFOI over period $[0, t]$ of the model (5.30) is

$$Y(t;0) = \frac{\beta P_l(t) R(t)}{\bar{\gamma}} + \hat{\mathcal{R}}_p(t) \tag{5.34}$$

Moreover, pluging $t = \infty$ into the above fomula (5.34) deduces the FSR of the model (5.30) to

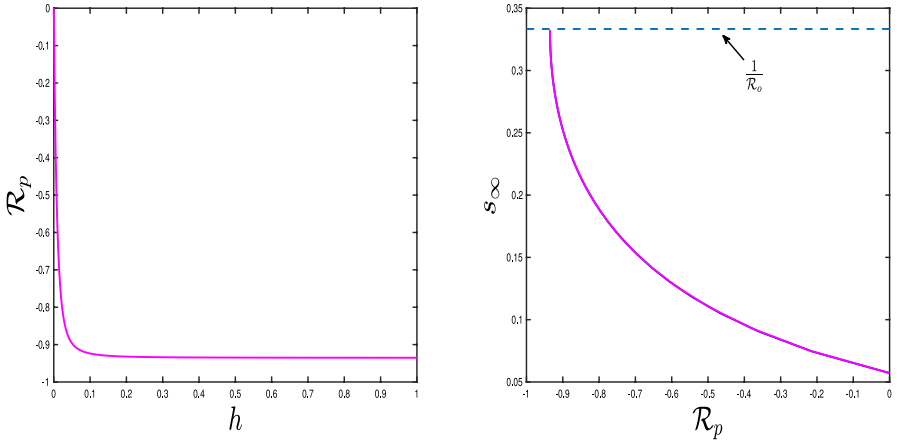

Fig. 4 $\mathcal{R}_o = 3, \bar{\gamma} = \frac{1}{20}, \epsilon = \frac{1}{14}, h \in [0, 5]$ and $N_0 = 300, E_0 = 5, I_0 = 5, R(0) = 0$

$$\ln \frac{s_0}{s_\infty} = \mathcal{R}_o r(\infty) + \mathcal{R}_p.$$

which can be rewritten as

$$\ln s_\infty = \mathcal{R}_0 s_\infty + [\ln s_0 - \mathcal{R}_o - \mathcal{R}_p]$$

The equation clearly shows that s_∞ is decreasing in $\mathcal{R}_p$, implying the NPIs or behaviour changes reflected by the fraction function $P_l(t) = P(L(t))$ can increase the final size s_∞ provided that $\mathcal{R}_p$ is no less than the minimal (critical) value $\mathcal{R}_p^{cr}$, where

$$\mathcal{R}_p^{cr} = 1 + \ln s_0 - \mathcal{R}_0 + \ln \mathcal{R}_0.$$

The above conclusion is obtained for (5.30) by analyzing its CFOI; similar results have also been discussed in Theorems 5.1 and 4.3 for other cases of the model. In summary, we have the following theorem.

Theorem 5.4 (See Fig. 4b) For any fixed $\mathcal{R}_o > 1$ is fixed, $\mathcal{R}_p$ decrease if and only if s_∞ increase (i.e. r_∞ decrease).

We point out that, by its definition, $\mathcal{R}_p$ (hence s_∞) actually depends on several things, including $\mathcal{R}_0$ through the kernel $\phi(a)$ and $P_l(t) = P(L(t))$. However, the dependence is generally very difficult to analyze. In Figs. 4 and 5, we present some numerical results to demonstrate such dependence. To this end, we choose $L(t) = I(t)$ and $P(L) = e^{-hL} = e^{-hI}$. Figure 4a shows the relationships between h and $\mathcal{R}_p$. By altering the non-biological parameter h, one can adjust the efficacy of NPIs (i.e. the value of $\mathcal{R}_p$). Figure 5 focuses the impact of the basic reproduction number $\mathcal{R}_o$ on both $\mathcal{R}_p$ and s_∞. Figure 5b is parallel and also is similar to Fig. 3a, and the explana-

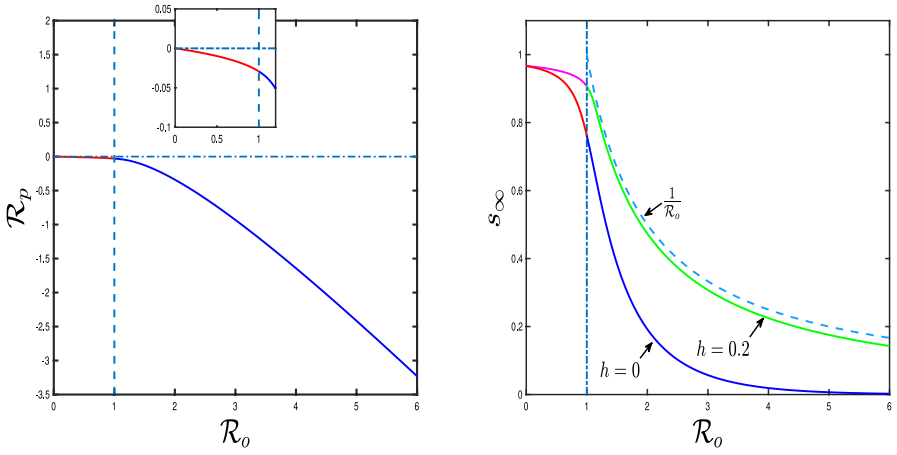

Fig. 5 $\mathcal{R}_o \in [0, 6], \bar{\gamma} = \frac{1}{20}, \epsilon = \frac{1}{14}, h = 0.2$ and $N_0 = 300, E_0 = 5, I_0 = 5, R(0) = 0$

tion of Fig. 5b is omitted here. In Theorem 5.2 and Remark 5.4, the theoretical results for the peak epidemic size of the case $L = U$ are presented. In practice, the number of the exposed class E is sometimes more difficult to collect precisely, whereas the mathematical analysis of the case $L = I$ is relatively complicated.

Figure 6 aims to present some numeric simulations on how the final size and peak size are impacted by the choice of $L(t)$ and h in the fraction function $P(L) = e^{-hL}$. The comparison is between $L = I$ and $L = U$ with different values of h. The numeric results seem to suggest that, in controlling peak size and final size, the effectiveness of Non-pharmaceutical and Non-biological precautions when applying only the number of $I(t)$ to measure the severity level $L(t)$ is not as good as when considering all infected numbers (i.e. $U(t)$), but it is still better than no precaution (i.e. $h = 0$).

6 Discussion

In this paper, we aim to explore the final epidemic size for general epidemics characterized by infection age, focusing on the effect of behaviour changes on the epidemic size. We employ the renewal equation approach, the power of which seems to be under-estimated in the literature. Our results show that the final size relation depends on both $\mathcal{R}_o$ and $\mathcal{R}_p$ (or $\mathcal{R}_{op}$) with $\mathcal{R}_p$ (or $\mathcal{R}_{op}$) accounting for the effect of behaviour changes. In addition, by the approach in Breda et al. (2012), we figure out the relation of the final size to the CFOI. As described in Remark 4.1, behaviour change alters

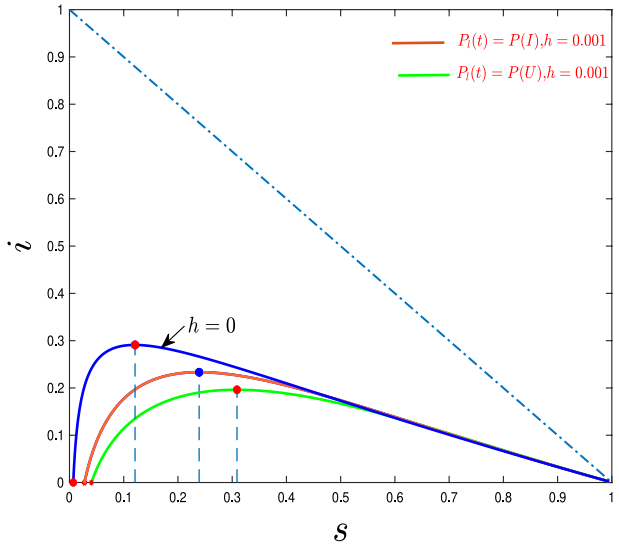

Fig. 6 $\mathcal{R}_o = 5, \bar{\gamma} = \frac{1}{20}, \epsilon = \frac{1}{14}$ and $N_0 = 2000, E_0 = 5, I_0 = 5, R(0) = 0$. Here, $h = 0$ represents the scenario without behavioural adaptation, where h serves as the flexibility index with which the population adaptively responds to the severity of epidemics. A larger h means stronger behaviour response. The extreme points of the curves represent the peak size, i_{max}, while the x-axis value at these extreme points corresponds to the herd immunity threshold, $s(t_m)$. Additionally, the leftmost zero points, which lie to the left of $x = s(t_m)$, correspond to s_∞, or equivalently, $1 - r_\infty$

this relation via introducing the parameter $\mathcal{R}_p$, in contrast to the results in Breda et al. (2012).

The assumption that time extends along the entire axis ($t_0 = -\infty$) simplifies mathematical analysis but makes numerical simulations impractical. By adopting this extension, some researchers are able to establish relatively more theoretical conclusions, but they can obtain fewer quantitative simulating results (Breda et al. 2012; Brauer 2008; Diekmann and Inaba 2023). Limiting the time starting from a finite value ($t_0 > -\infty$), we repeat, in Sect. 4.2, the derivations and inductions in great detail to overcome the stumbling block in numerical simulations. Section 4.2 places more emphasis on the contribution of initial states than Sect. 4.1, through incorporating the term $\hat{\mathcal{R}}_g(t)$ or $\mathcal{R}_g$.

The theorems on the FSR in Sect. 4 provide theoretical support for estimating the magnitude of the epidemic through the estimate of the basic reproduction number $\mathcal{R}_o$ as well as the intervention effect term $\mathcal{R}_p$ which is not easy to determine. In Sect. 5, by choosing two particular and simple forms for kernel function $\varphi(a)$, we are led to a differential equation for $\hat{\mathcal{R}}_p(t)$ or $\hat{\mathcal{R}}_{op}(t)$ [see the last equation of (5.10) or 5.19], which can be analyzed and simulated easily. Further research is needed to analyze $\mathcal{R}_p$ so that more information can be gathered on the impact of fraction function $P_l(t) = P(L(t))$ on the final size.

As mentioned in the previous sections, the Conjecture 4.1 is analytically challenging, although its conclusion is essential and biologically meaningful. It is certainly worth working on. Thus, an immediate research project is to establish sufficient conditions under which the Conjecture 4.1 holds.

Section 5 shows that for some special cases as dealt with in Sects. 5.20 and 5.2, the framework of the general infection-age model (3.14) reduced to simpler forms that can actually offer some information about the herd immunity threshold (HIT), the peak size (PZ) and the calendar time t_m. In this Section, we partially answer the **(Q1)** posed at the beginning, both theoretically and numerically, showing that behavioural adaption can effectively reduce (resp. increase) the final size r_∞ (resp. increase s_∞), PZ and HIT in concrete cases (see for instance Figs. 5 and 8 along with their descriptions). However, for the general case of the general model (3.14), it is very challenging to analyze the HIT and PZ. More mathematical techniques are needed to theoretically obtain information on the HIT and PZ for the general model (3.14).

As introduced earlier, following (Cheng and Zou 2022, 2024), we adopt the notion of "practically susceptible population" and consider *susceptibility* at the population level. In fact, more reasonably, adaptive behaviour changes are heterogeneous due to different cultures, societal norms, economic costs, etc. Some recent work (Almeida et al. 2021; Berestycki et al. 2023; Diekmann and Inaba 2023; Tkachenko et al. 2021), and references therein take into account heterogeneity in both infection risk and social mechanisms. For example, Diekmann and Inaba (2023) incorporate separable static heterogeneity into compartmental models using RE methods with less focus on the final size equation/relation. In reality, humans' behaviour response to infection risk exhibits *adaptive/or dynamic* heterogeneity, which compartmental models do not accurately capture. Additionally, the RE method is more powerful for generalization than compartmental models. The theoretical derivations of the final

size relation by RE-methods, coupled with individual-level heterogeneity of adaptive behaviour change, are often mathematically challenging but meaningful.

In summary, we believe this work first provides a general and plausible framework for estimating the final epidemic size when considering Non-pharmaceutical interventions and/or psychological effects. From a mathematical point of view, this work generalizes the force of infection described in Cheng and Zou (2022) to reflect the above-mentioned effects and derives the corresponding REs. As is seen, the incorporation of Non-pharmaceutical interventions and/or psychological effects results in REs with *time-varying kernels*. It is the time-varying nature of the kernel that makes the model novel, the analysis more difficult, and the results more interesting both mathematically and biologically.

Appendix

The derivation of model (3.14)

$$\begin{aligned} E'(t) = \int_0^\tau \frac{\partial u(t,a)}{\partial t} da &= \int_0^\tau -\gamma(a)u(t,a) - \frac{\partial u(t,a)}{\partial a} da \\ &= -\int_0^\tau \gamma(a)u(t,a)da - u(t,\tau) + u(t,0) \\ &= -\int_0^\tau \gamma(a)u(t,a)da - u(t,\tau) + B(t) \\ I'(t) = \int_\tau^\infty \frac{\partial u(t,a)}{\partial t} da &= -\int_\tau^\infty \gamma(a)u(t,a)da - \int_\tau^\infty \frac{\partial u(t,a)}{\partial a} da \\ &= -\int_\tau^\infty \gamma(a)u(t,a)da - u(t,\infty) + u(t,\tau) \\ &= -\int_\tau^\infty \gamma(a)u(t,a)da + u(t,\tau) \end{aligned} \tag{.1}$$

Proof of Theorem 3.1

$$\frac{d(S+U)}{dt} = -\int_0^\infty \gamma(a)u(t,a)da = -\int_0^\infty B(t-a)\sigma(a)\gamma(a)da \tag{.2}$$

which is strictly negative, $S+U$ is strictly decreasing and positive. Therefore, $\lim_{t\to\infty} \frac{d(S+U)}{dt} = 0$, that is, $U(\infty) = 0$. Further, since $I(t), E(t) > 0$ for all $t < \infty$, $E(\infty) = 0, I(\infty) = 0$. □

Proof of Lemma 4.3 Under the Assumption 4.2 of $u_0(a)$, then

$$\min\left\{\frac{1}{V_\tau}, 1\right\} G_M(t) \le G(t) \le \max\left\{\frac{1}{V_\tau}, 1\right\} G_M(t)$$

where

$$G_M(t) = (E_0 + I_0)\varphi(t)\frac{\int_0^\infty u_0(\eta)\frac{\beta(\eta+t)}{\beta(t)}d\eta}{\int_0^\infty u_0(\eta)d\eta}, \quad t \geq \tau. \tag{.3}$$

Moreover, particularly, if $\beta(a) \equiv \beta$,

$$\int_\tau^t P_l(a)G_M(a)da = \frac{E_0 + I_0}{N_0}\hat{\mathcal{R}}_c(t), \tag{.4}$$

$$\min\left\{\frac{1}{V_\tau}, 1\right\}\frac{N_0 - S_0}{N_0}\hat{\mathcal{R}}_c(t) \leq \hat{\mathcal{R}}_g(t) \leq \max\left\{\frac{1}{V_\tau}, 1\right\}\frac{N_0 - S_0}{N_0}\hat{\mathcal{R}}_c(t) \tag{.5}$$

and

$$\begin{aligned}
&\min\left\{\frac{1}{V_\tau}, 1\right\}\hat{\mathcal{R}}_c(t) \leq \left(\min\left\{\frac{1}{V_\tau}, 1\right\}\frac{E_0 + I_0}{N_0} + \frac{S_0}{N_0}\right)\hat{\mathcal{R}}_c(t) \leq \hat{\mathcal{R}}_{op}(t) - \hat{\mathcal{R}}_p(t) \\
&\leq \left(\max\left\{\frac{1}{V_\tau}, 1\right\}\frac{E_0 + I_0}{N_0} + \frac{S_0}{N_0}\right)\hat{\mathcal{R}}_c(t) \leq \max\left\{\frac{1}{V_\tau}, 1\right\}\hat{\mathcal{R}}_c(t) \\
&\leq \max\left\{\frac{1}{V_\tau}, 1\right\}\hat{\mathcal{R}}_o(t).
\end{aligned} \tag{.6}$$

This inequality implies (4.47). □

Proof of Lemma 4.2 For any $a < b$, with the help of the second mean value theorem for integrals (Roman et al. 2012; Dixon 1929), there exists $\xi_{a,b} \in [a, b]$, which depend on the choice of a, b, such that

$$\begin{aligned}
\int_a^b g(t)f'(t)dt &= g(a)\int_a^{\xi_{a,b}} f'(t)dt + g(b)\int_{\xi_{a,b}}^b f'(t)dt \\
&= g(a)\left(f(\xi_{a,b}) - f(a)\right) + g(b)\left(f(b) - f(\xi_{a,b})\right).
\end{aligned} \tag{.7}$$

Further, we get the following inequalities:

$$\begin{aligned}
&g(a)\left(f(\xi_{a,b}) - f(a)\right) + g(b)\left(f(b) - f(\xi_{a,b})\right) \\
&< g(a)\left(f(\xi_{a,b}) - f(a)\right) + g(b)\left(f_M - f(\xi_{a,b})\right) \\
&< g_M\left(f_M - f(a)\right) + g_M\left(f_M - f(\xi_{a,b})\right) \\
&< g_M\left(f_M - f(a)\right) < g_M f_M
\end{aligned} \tag{.8}$$

and

$$\begin{aligned} & g(a)\left(f(\xi_{a,b})-f(a)\right)+g(b)\left(f(b)-f(\xi_{a,b})\right) \\ & \quad = f(\xi_{a,b})\left(g(a)-g(b)\right)+g(b)f(b)-g(a)f(a) \\ & \quad > f_M(g_m-g_M)+g(b)f(b)-g(a)f(a) \\ & \quad > f_M(g_m-g_M)-g(a)f(a) \\ & \quad > f_M(g_m-g_M)-g_Mf_M = f_M(g_m-2g_M) \end{aligned} \tag{.9}$$

Let $a \to -\infty$ and $b \to \infty$,

$$f_M(g_m-2g_M) < \int_{-\infty}^{+\infty} g(t)f'(t)dt < g_Mf_M.$$

Moreover, if $g : \mathbb{R} \to (g_m, g_M)$ is strictly decreasing,

$$\begin{aligned} & g(a)\left(f(\xi_{a,b})-f(a)\right)+g(b)\left(f(b)-f(\xi_{a,b})\right) \\ & \quad = f(\xi_{a,b})\left(g(a)-g(b)\right)+g(b)f(b)-g(a)f(a) \\ & \quad > g(b)f(b)-g(a)f(a) \end{aligned} \tag{.10}$$

holds. Hence,

$$(g_m-g_M)f_M \le \int_{-\infty}^{+\infty} g(t)f'(t)dt.$$

Choose $a = -b < 0 < b$ and there exists $\xi_b \in [-b, b]$ such that

$$\begin{aligned} & \int_{-b}^{b} g(t)f'(t)dt \\ & \quad \le g(-b)\left(f(\xi_b)-f(-b)\right)+g(-b)\left(f_M-f(\xi_b)\right) \\ & \quad = g(-b)\left(f_M-f(-b)\right) \end{aligned} \tag{.11}$$

Let $b \to +\infty$,

$$\int_{-\infty}^{\infty} g(t)f'(t)dt \le 0. \tag{.12}$$

□

Proof of Lemma 4.3 If choosing another set of parameters and keeping $\mathcal{R}_o$ the same, we get

$$\begin{aligned} -\Delta\mathcal{R}_p &= \ln(s_\infty+\Delta s_\infty)+\mathcal{R}_o(1-(s_\infty+\Delta s_\infty))+\mathcal{R}_p \\ &= \ln\left(1+\frac{\Delta s_\infty}{s_\infty}\right)-\mathcal{R}_o\Delta s_\infty \end{aligned} \tag{.13}$$

Moreover, if $\frac{\Delta s_\infty}{s_\infty}$ is small, we find by straightforward Taylor expansion that

$$\begin{aligned}\Delta\mathcal{R}_p &= -\left(\frac{\Delta s_\infty}{s_\infty} + \mathcal{O}\left(\frac{\Delta s_\infty}{s_\infty}\right) - \mathcal{R}_o \Delta s_\infty\right) \\ &\approx \left(\mathcal{R}_o - \frac{1}{s_\infty}\right)\Delta s_\infty\end{aligned} \tag{.14}$$

or

$$\lim_{\Delta s_\infty \to 0} \frac{\Delta\mathcal{R}_p}{\Delta s_\infty} = \mathcal{R}_o - \frac{1}{s_\infty} \tag{.15}$$

□

Proof of Lemma 5.1

$$\begin{aligned}\int_0^b f'(t)g(t)dt &= g(0)\int_0^{\xi_b} f'(t)dt + g(b)\int_{\xi_b}^b f'(t)dt \\ &= g(0)\left[f(\xi_b) - f(0)\right] + g(b)[f(b) - f(\xi_b)] \\ &= g(b)[f(b) - f(\xi_b)]\end{aligned} \tag{.16}$$

$$g(b)f(b) > \int_0^b f'(t)g(t)dt > g(b)f(b) - f_M g(b) \tag{.17}$$

Let $b \to \infty$,

$$g_M f_M \geq \int_0^\infty f'(t)g(t)dt \geq g_M f_M - f_M g_M = 0. \tag{.18}$$

□

Proof of the Theorem 5.1 Suppose to the contrary $s_\infty > \frac{1}{\mathcal{R}_o}$, which indicates that there exists a $\delta > 0$ such that $s_\infty = \frac{1}{\mathcal{R}_o} + \delta$ and then $s(t) \geq \frac{1}{\mathcal{R}_o} + \delta$ for all $t \geq 0$ owing to the decrement of $s(t)$. Since $\lim_{t\to\infty} s_P(t) = P_l(\infty)s_\infty = P(0)s_\infty$ and $\delta > 0$, for given $(P(0)\delta)/2$ there exists $\bar{t}_0 \in [0, \infty)$ such satisfying

$$|s_P(t) - P(0)s_\infty| < \frac{P(0)\delta}{2}$$

for all $t \geq \bar{t}_0$. Thus, we have

$$s_P(t) > \frac{P(0)}{\mathcal{R}_o} + \frac{P(0)\delta}{2}$$

and

$$U'(t) = \mathcal{R}_o \left(\frac{s_P(t)}{P(0)} - \frac{1}{\mathcal{R}_o} \right) \bar{\gamma} I > \frac{\mathcal{R}_o \bar{\gamma} I \delta}{2}$$

for all $t \geq \bar{t}_0$. Then, for all $t \geq \bar{t}_0$, we have

$$U(t) > U(t_0) e^{\frac{\mathcal{R}_o \bar{\gamma} \delta}{2} \int_{t_0}^{t} \frac{I(t)}{U(t)} dt} > U(t_0)$$

which means $U(\infty) = +\infty$, it contradicts with $U(\infty) = 0$. Thus, $s_\infty \leq \frac{1}{\mathcal{R}_o}$ holds.

Moreover, if $P_l(t) = P(U)$, $s_\infty < \frac{1}{\mathcal{R}_o}$. Suppose to the contrary $s_\infty = \frac{1}{\mathcal{R}_o}$ i.e. $S_\infty = \frac{N_0}{\mathcal{R}_o}$. Then

$$\begin{aligned} \frac{dU}{dS} &= -\frac{\mathcal{R}_o \left(\frac{s_P(t)}{P(0)} - \frac{1}{\mathcal{R}_o} \right) \bar{\gamma}}{\beta N_0 s_P(t)} = -\frac{\left(s_P(t) - \frac{P(0)}{\mathcal{R}_o} \right)}{s_P(t)} \\ \frac{d}{dS}\frac{dU}{dS} &= -\frac{d}{ds_P} \left[\frac{\left(s_P(t) - \frac{P(0)}{\mathcal{R}_o} \right)}{s_P(t)} \right] \frac{ds_P}{dS} \\ &= -\frac{P(0)}{\mathcal{R}_o} \frac{1}{s_P^2} \frac{ds_P}{dS} = -\frac{P(0)}{\mathcal{R}_o} \frac{1}{s_P^2} \left[\frac{dP_l}{dU} \frac{dU}{dS} s + \frac{P_l}{N_0} \right] \end{aligned} \tag{.19}$$

which means $\frac{d^2U}{dS^2} < 0$ if $\frac{dU}{dS} \leq 0$.

$$\left. \frac{dU}{dS} \right|_{S=S_\infty} = \frac{\left(P(U(S_\infty)) s_\infty - \frac{P(0)}{\mathcal{R}_o} \right)}{P(U(S_\infty)) s_\infty} = \frac{\mathcal{R}_o \left(\frac{P(U(S_\infty))}{\mathcal{R}_o} - \frac{P(0)}{\mathcal{R}_o} \right)}{P(U(S_\infty))} = 0.$$

Thus, there is a $\delta > 0$ such that for any $S(t) \in [S_\infty, S_\infty + \delta)$, $\frac{dU}{dS} < 0$. And hence, $U(S_\infty) = 0 > U(S_\infty + \delta/2)$, a contradiction. □

Proof of Lemma 5.3 Since $U(t) > 0$ for all $t \in [0, \infty)$, $U'(t_m) = 0$ tells

$$\frac{s_P(t_m)}{P(0)} - \frac{1}{\mathcal{R}_o} = 0.$$

Then, we have

$$
\begin{aligned}
U''(t_m) &= \mathcal{R}_o \left(\frac{s'_P(t_m)}{P(0)} \right) \bar{\gamma} I(t_m) \\
&= \mathcal{R}_o \left(\frac{P'(U)U'(t_m)s(t_m) + P(U(t_m))s'(t_m)}{P(0)} \right) \bar{\gamma} I(t_m) \\
&= \mathcal{R}_o \left(\frac{P(U(t_m))s'(t_m)}{P(0)} \right) \bar{\gamma} I(t_m) \leq 0
\end{aligned} \tag{.20}
$$

□

The details in reduction of the model (5.30)

Inspired by Diekmann and Inaba (2023), we rewrite $\varphi_2(a)$ as a vector form

$$
\varphi_2(a) = \boldsymbol{\beta}(a)^T e^{\mathcal{B}a} \mathrm{Id}_{2\times 1} \tag{.21}
$$

where

$$
\beta(a) := \begin{pmatrix} 0 \\ \beta \end{pmatrix}, \mathrm{Id}_{2\times 1} := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \gamma(a) := \mathcal{B} = \begin{bmatrix} -\epsilon & 0 \\ \epsilon & -\bar{\gamma} \end{bmatrix}.
$$

The infected density $\mathrm{u}(t,a)$ on infection-age a corresponding to the kernel $\varphi(a) = \varphi_2(a)$ is a two-dimensional vector:

$$
\mathrm{u}(t,a) = \begin{pmatrix} u_e(t,a) \\ u_i(t,a) \end{pmatrix},
$$

which is different from the Eq. (2.1):

$$
\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial a} \right) \mathrm{u}(t,a) = \mathcal{B}\mathrm{u}(t,a) \tag{.22}
$$

with boundary condition $\mathrm{u}(t,0) = [u_e^0(t), 0]^T$. The age densities of total infected individuals $u(t,a) := ||\mathrm{u}(t,a)||_1 = u_e(t,a) + u_i(t,a)$. Further, denote

$$
E(t) = \int_0^{+\infty} u_e(t,a)da,\ I(t) = \int_0^{+\infty} u_i(t,a)da.
$$

exposed and infectious individuals, respectively. Further, simple calculations lead to the model (5.30)

The figure and results of model (5.20) in Subsect. 5.1.2

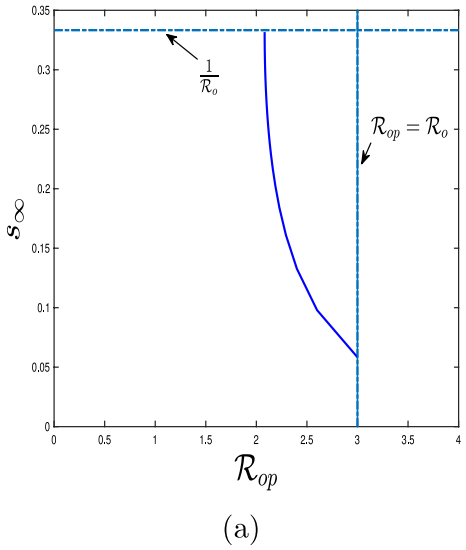

Fig. 7 $h \in [0, 2], \mathcal{R}_o = 3$. Other parameters are $\gamma_2 = \frac{1}{7}, N_0 = 300, E_0 = 5, I_0 = 5, R(0) = 0$

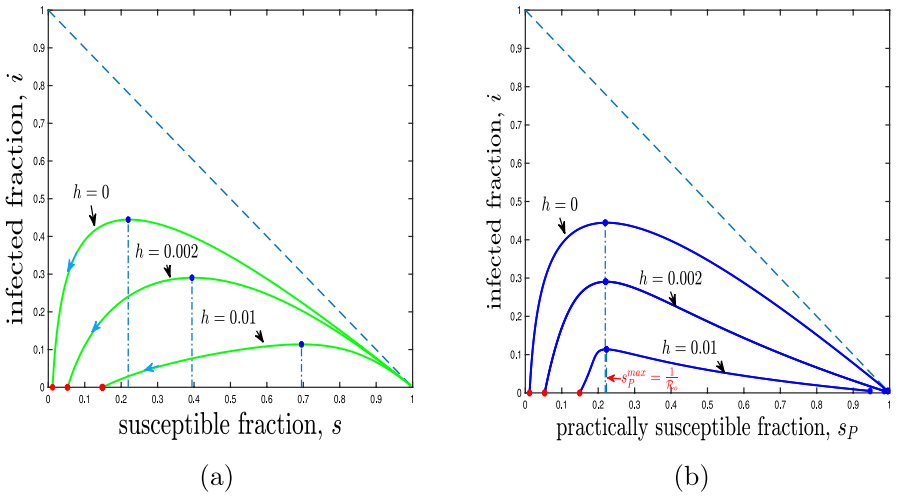

Fig. 8 Parameters are $\mathcal{R}_o = 4.5, \gamma_2 = \frac{1}{7}, N_0 = 1000, I_0 = 5, R(0) = 0$. The two figures demonstrate that as h increases, the peak size reduces sequentially, the *threshold for herd immunity(HIT)* increases sequentially (see the extreme points of these three curves), and the final size increases (see the three leftmost zero points on the x-axis). Additionally, regardless of changes in h, the threshold for *practically susceptible* fraction s_P^{max} remains at $1/\mathcal{R}_o$, which actually is the classical herd immunity level without behavioural adaption

Acknowledgements The authors are very grateful to the two anonymous reviewers for their careful reading and valuable comments which have led to significant improvement in the presentation of the paper.

Data availability This work is mainly of a theoretical and analytical nature, and hence, there is no data involved.

Declarations

Conflict of interest This work has no conflict of interest.

References

Almeida L, Bliman PA, Nadin G, Perthame B, Vauchelet N (2021) Final size and convergence rate for an epidemic in heterogeneous populations. Math Models Methods Appl Sci 31(05):1021–1051
Andreasen V (2011) The final size of an epidemic and its relation to the basic reproduction number. Bull Math Biol 73(10):2305–2321
Anderson D, Watson R (1980) On the spread of a disease with gamma-distributed latent and infectious periods. Biometrika 67(1):191–198
Anderson RM, May RM (1991) Infectious diseases of humans: dynamics and control. Oxford University Press, Oxford
Arino J, Brauer F, van den Driessche P, Watmough J, Wu J (2007) A final size relation for epidemic models. Math Biosci Eng 4(2):159
Berestycki H, Desjardins B, Weitz JS, Oury JM (2023) Epidemic modeling with heterogeneity and social diffusion. J Math Biol 86(4):60
Brauer F (2005) The Kermack–McKendrick epidemic model revisited. Math Biosci 198(2):119–131
Brauer F (2008) Age of infection and the final size relation. Math Biosci Eng 5(4):681
Breda D, Diekmann O, De Graaf WF, Pugliese A, Vermiglio R (2012) On the formulation of epidemic models (an appraisal of Kermack and McKendrick). J Biol Dyn 6(sup2):103–117
Champredon D, Dushoff J, Earn DJ (2018) Equivalence of the Erlang-distributed SEIR epidemic model and the renewal equation. SIAM J Appl Math 78(6):3258–3278
Cheng T, Zou X (2022) A new perspective on infection forces with the demonstration by a DDE infectious disease model. Math Biosci Eng 19(5):4856–4880
Cheng T, Zou X (2024) Modelling the impact of precaution on disease dynamics and its evolution. J Math Biol 89(1):1
Diekmann O, Heesterbeek JAP (2000) Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. Wiley, Hoboken
Diekmann O, Inaba H (2023) A systematic procedure for incorporating separable static heterogeneity into compartmental epidemic models. J Math Biol 86(2):1–19
Diekmann O, Hans H, Tom B (2013) Mathematical tools for understanding infectious disease dynamics, vol 7. Princeton University Press, Princeton
Diekmann O, Othmer HG, Planque R, Bootsma MC (2021) The discrete-time Kermack–McKendrick model: a versatile and computationally attractive framework for modelling epidemics. Proc Natl Acad Sci USA 118(39):e2106332118
Dixon AC (1929) The second mean value theorem in the integral calculus. Math Proc Camb Philos Soc 25(3):282–284
Feller W (1940) On the logistic law of growth and its empirical verifications in biology. Acta Biotheor 5:51–66
Feller W (1941) On the integral equation of renewal theory. Ann Math Stat 12(3):243–267
Feng Z (2007) Final and peak epidemic sizes for SEIR models with quarantine and isolation. Math Biosci Eng 4(4):675
Fine P, Eames K, Heymann DL (2011) "Herd immunity": a rough guide. Clin Infect Dis 52(7):911–916
Heng K, Althaus CL (2020) The approximately universal shapes of epidemic curves in the Susceptible-Exposed-Infectious-Recovered (SEIR) model. Sci Rep 10(1):1–6
Inaba H (2017) Age-structured population dynamics in demography and epidemiology. Springer, Berlin

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A Contain Pap Math Phys Charact 115(772):700–721
Ketcheson DI (2021) Optimal control of an SIR epidemic through finite-time non-pharmaceutical intervention. J Math Biol 83(1):7
Kröger M, Turkyilmazoglu M, Schlickeiser R (2021) Explicit formulae for the peak time of an epidemic from the SIR model. Which approximant to use? Phys D Nonlinear Phenom 425:132981
Lebreton JD (1996) Demographic models for subdivided populations: the renewal equation approach. Theor Popul Biol 49(3):291–313
Lotka AJ (1907) Relation between birth rates and death rates. Science 26(653):21–22
Ma J, Earn DJ (2006) Generality of the final size formula for an epidemic of a newly invading infectious disease. Bull Math Biol 68:679–702
Magal P, McCluskey CC, Webb GF (2010) Lyapunov functional and global asymptotic stability for an infection-age model. J Appl Anal 89(7):1109–1140
Magal P, Seydi O, Webb G (2016) Final size of an epidemic for a two-group SIR model. SIAM J Appl Math 76(5):2042–2059
Miller JC (2012) A note on the derivation of epidemic final sizes. Bull Math Biol 74(9):2125–2141
Roman W, Hetmaniok E, Damian S (2012) A stronger version of the second mean value theorem for integrals. Comput Math Appl 64(6):1612–1615
Sereno J, Anderson A, Ferramosca A, Hernandez-Vargas EA, González AH (2022) Minimizing the epidemic final size while containing the infected peak prevalence in SIR systems. Automatica 144:110496
Sharpe FR, Lotka AJ, Alfred JL (1911) A problem in age-distribution. Lond Edinb Dublin Philos Mag J Sci 21(124):435–438
Tkachenko AV, Maslov S, Wang T, Elbana A, Wong GN, Goldenfeld N (2021) Stochastic social behavior coupled to COVID-19 dynamics leads to waves, plateaus, and an endemic state. Elife 10:e68341