Journal of Mathematical Biology (2023) 86:37
https://doi.org/10.1007/s00285-023-01869-w

Mathematical Biology

Viral dynamics with immune responses: effects of distributed delays and Filippov antiretroviral therapy

Jiawei Deng[1] · Hongying Shu[1] · Lin Wang[2] · Xiang-Sheng Wang[3]

Received: 15 July 2022 / Revised: 17 October 2022 / Accepted: 6 January 2023 /
Published online: 25 January 2023

Abstract

In this paper, we propose a general viral infection model to incorporate two infection modes (virus-to-cell mode and cell-to-cell mode), the CTL immune response, and the distributed intracellular delays during the processes of viral infection, viral production, and CTLs recruitment. We investigate the existence, the uniqueness, and the global stability of three equilibria: infection-free equilibrium E_0, immune-inactivated equilibrium E_1 and immune-activated equilibrium E_2, respectively. We prove that the viral dynamics are determined by two threshold parameters: the basic reproduction number for infection R_0 and the basic reproduction number for immune response R_{IM}. We also numerically explore the viral dynamics beyond stability. We use bifurcation diagrams to show that increasing the delay in CTL immune cell recruitment can induce a switch in viral load from a stable constant level to sustained oscillations, and then back to a stable equilibrium. We also compare the contributions of the two infection modes to the total infection level and identify the key parameters that would affect the percentages of virus-to-cell infection and cell-to-cell infection. Finally, we explore how Filippov control can be applied in antiretroviral therapy to reduce the viral loads.

Keywords Viral dynamics with immune responses · Cell-to-cell infection · Virus-to-cell infection · Distributed delays · Filippov antiretroviral therapy

Mathematics Subject Classification 92D30 · 37N25 · 34K20

✉ Hongying Shu
hshu@snnu.edu.cn

1 School of Mathematics and Statistics, Shaanxi Normal University, Xi'an 710062, People's Republic of China

2 Department of Mathematics and Statistics, University of New Brunswick, Fredericton, NB E3B 5A3, Canada

3 Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70503, USA

1 Introduction

Mathematical models have been used to understand the in vivo infection dynamics of many viruses (Bonhoeffer et al. 1997; Perelson et al. 1993; Perelson and Nelson 1999; Perelson et al. 1996). These in-host models were applied: (i) to describe the viral infection process (Perelson et al. 1993; Perelson and Nelson 1999); (ii) to estimate key parameters such as the virion clearance rate and the life span of infected cells (Perelson et al. 1996); and (iii) to guide the development of antiviral drug therapies (Bonhoeffer et al. 1997). The viral dynamics can be described by the differential equations of three compartments: the uninfected target cells, the infected cells, and the free virus; sometimes there is a fourth compartment corresponding to the immune response (Chen et al. 2016; Nowak and Bangham 1996; Wang et al. 2013). The immune cells such as the cytotoxic T-lymphocyte cells (CTLs) play a crucial role in the defense of viral infections. A mathematical model with CTL immune response was proposed in (Nowak and Bangham 1996) to explore the dynamics of kinetic interaction between CTLs and infected cells. Two HIV models were further developed in (Wang et al. 2013) to incorporate a logistic growth term for the uninfected target cells.

Many of the existing mathematical models focused only on the virus-to-cell infection mode; see for example, (Leenheer and Smith 2003; Perelson and Nelson 1999) for human immunodeficiency virus (HIV), (Ribeiro et al. 2002; Tsiang et al. 1999) for hepatitis B virus (HBV), and (Dahari et al. 2007; Neumann et al. 1998) for hepatitis C virus (HCV). However, another mode called the cell-to-cell infection mode has also been recognized in clinics (Sato et al. 1992; Sourisseau et al. 2007). It is the process when the viral particles are transferred directly from an infected cell to an uninfected target cell through the formation of virological synapses (Galloway et al. 2015; Hübner et al. 2009). For some viruses, the cell-to-cell infection mode seems to be more potent and efficient than the virus-to-cell infection mode (Komarova et al. 2013a, b), and may contribute to more than half of the viral infections (Iwami et al. 2015). Even with antiretroviral therapy, the cell-to-cell spread of HIV can still permit ongoing replication (Sigal et al. 2011). It is thus important to consider both virus-to-cell and cell-to-cell infection modes in viral infection models (Martin and Sattentau 2009). Some earlier studies of the cell-to-cell infection can be found in (Gummuluru et al. 2000; Dixit and Perelson 2004).

Denote by $T(t)$, $I(t)$, $V(t)$ and $Z(t)$ the concentrations of uninfected target cells, infected cells, viruses, and virus-specific CTLs at time t, respectively. A general viral infection model with two infection modes and the CTL immune response is given as below.

$$\begin{aligned}
T'(t) &= b(T(t)) - g(T(t), I(t)) - h(T(t), V(t)),\\
I'(t) &= g(T(t), I(t)) + h(T(t), V(t)) - \mu_1 I(t) - pI(t)Z(t),\\
V'(t) &= kI(t) - \mu_2 V(t),\\
Z'(t) &= qI(t)Z(t) - \mu_3 Z(t),
\end{aligned}$$

where $b(T)$ denotes the intrinsic growth rate of the uninfected target cells, which includes both production and natural mortality. The nonlinear functions $g(T, I)$ and $h(T, V)$ are the incidence rates of new infections through cell-to-cell and virus-to-cell transmissions, respectively. The infected cells produce virions at a rate kI, die at a per

capita rate μ_1, and are cleared by CTLs at a rate of pIZ. The virions are cleared at a per capita rate μ_2. The CTLs are recruited at a rate qIZ, and decay at a per capita rate μ_3.

Time delays in the processes of viral infection and replication have been shown to play a significant role in understanding the viral dynamics (Herz et al. 1996; Huang et al. 2010; Li and Shu 2010). The infected cells would go through a latent period before they burst virions. The released virions may need a maturation phase outside the cell before becoming infectious. Moreover, there is a complex chain of events between the CTL attack and the subsequent recruitment (Canabarro et al. 2004; Wang et al. 2007). Thus, it is important to introduce time delays in each of the aforementioned steps. The delay during the production of an actively infected target cell through either cell-to-cell or virus-to-cell transmission is modeled by the density function $f_1(\tau)$, which describes the probability that an infected cell survives and becomes actively infectious τ time units past the infection. Following the modeling ideas in (Nakata 2010; Shu et al. 2018), we assume the production rate of new actively infected cells at time t to be

$$\int_0^\infty f_1(\tau)(g(T(t-\tau), I(t-\tau)) + h(T(t-\tau), V(t-\tau)))d\tau.$$

We use another density function $f_2(\tau)$ to describe the survival and maturation probability of free viral particles τ time units after being released. The rate of virion maturation at time t is $k\int_0^\infty f_2(\tau)I(t-\tau)d\tau$. Finally, we use $f_3(\tau)$ to describe the distribution of delays between cell encounters and subsequent recruitment. The rate of CTLs proliferation at time t is then given by $q\int_0^\infty f_3(\tau)I(t-\tau)Z(t-\tau)d\tau$. The general viral infection model incorporating two infection modes, CTL immune response and distributed delays can be described by the following system of differential-integral equations:

$$\begin{aligned}
T'(t) &= b(T(t)) - g(T(t), I(t)) - h(T(t), V(t)),\\
I'(t) &= \int_0^\infty f_1(\tau)(g(T(t-\tau), I(t-\tau)) + h(T(t-\tau), V(t-\tau)))d\tau\\
&\quad - \mu_1 I(t) - pI(t)Z(t),\\
V'(t) &= k\int_0^\infty f_2(\tau)I(t-\tau)d\tau - \mu_2 V(t),\\
Z'(t) &= q\int_0^\infty f_3(\tau)I(t-\tau)Z(t-\tau)d\tau - \mu_3 Z(t),
\end{aligned} \tag{1.1}$$

where as before $T(t)$, $I(t)$, $V(t)$ and $Z(t)$ are the concentrations of uninfected target cells, actively infected target cells, mature viruses, and virus-specific CTLs at time t, respectively. All parameters in (1.1) are positive. The density functions $f_i(\tau)$ (with $i = 1, 2, 3$) satisfy the conditions

$$f_i(\tau) \geq 0 \text{ for all } \tau \geq 0, \quad \int_0^\infty f_i(\tau)d\tau > 0,$$
$$\text{and} \quad \int_0^\infty f_i(\tau)e^{s\tau}d\tau < \infty \text{ for some positive } s.$$

Throughout this paper, we assume that the intrinsic growth function of uninfected target cells b, and two incidence rates of infections g and h satisfy the following conditions.

($\mathbf{H}_1$) $b(T) \in C^1(\mathbb{R}_+)$, there exists $\overline{T} > 0$ such that $b(\overline{T}) = 0$, and $b(T)(T - \overline{T}) < 0$ for $T \neq \overline{T}$ and $T > 0$.

($\mathbf{H}_2$) $g(u, v), h(u, v) \in C^{1,2}(\mathbb{R}_+ \times \mathbb{R}_+)$ are strictly increasing with respect to both variables, concave down with respect to the second variable v, and vanish if and only if $uv = 0$.

Here, the constant $\overline{T}$ is biologically interpreted as the natural level of target cells in the absence of viral infections. Some typical examples of $b(T)$ satisfying ($\mathbf{H}_1$) are $\alpha - dT$, $\alpha - dT + rT(1 - T/T_m)$ and $\alpha + rT(1 - T/T_m)(T + a)$, where $\alpha, d, r, T_m, a > 0$. The examples of $g(u, v)$ and $h(u, v)$ satisfying ($\mathbf{H}_2$) include βuv, $\beta u^l v$, $\beta u^l v^2$ and $\beta uv/[(u + a_1)(v + a_2)]$, where $\beta, l > 0$ and $a_1, a_2 \geq 0$. Our general model (1.1) includes many existing models as special cases (Lai and Zou 2014; Nakata 2010; Wang et al. 2016).

The rest of this paper is organized as follows. In Sect. 2, we derive the basic reproduction number for infection and the basic reproduction number for immune response. In Sect. 3, we prove that these two basic reproduction numbers are threshold parameters in determining the global dynamics of our model. In Sect. 4, we numerically explore how the infection modes, the delay in immune cell recruitment, and the antiretroviral therapy with a Filippov control affect the viral dynamics. A summary and a discussion are presented in Sect. 5.

2 Preliminary results and the basic reproduction numbers

First, we show that system (1.1) is well-posed on the phase space $\mathcal{C}^4$, where $\mathcal{C}$ is the Banach space of fading memory type (Hale and Kato 1978)

$$\mathcal{C} = \Big\{\phi \in C\left((-\infty, 0], \mathbb{R}\right) \Big| \phi(\theta)e^{s\theta} \text{ is uniformly continuous for } \theta \in (-\infty, 0] \text{ and } \|\phi\| < \infty\Big\},$$

equipped with the norm $\|\phi\| = \sup_{\theta \leq 0} |\phi(\theta)|e^{s\theta}$ for a fixed constant $s > 0$. If ϕ is defined on the whole real line, then for each $t \geq 0$, we use the standard notation ϕ_t to denote a function defined by $\phi_t(\theta) = \phi(t + \theta)$ for $\theta \in (-\infty, 0]$. For biological applications, we assume that the initial condition of (1.1) is given in $\mathcal{C}_+^4$, where $\mathcal{C}_+ = \mathcal{C} \cap C((-\infty, 0], \mathbb{R}_+)$ is the nonnegative cone of $\mathcal{C}$. The existence and uniqueness of the solution of system (1.1) then follow from the theory of functional differential equations with infinitely distributed delays (Hale and Kato 1978). Now, we denote

$$\beta_i = \int_0^\infty f_i(\tau)d\tau, i = 1, 2, 3, \ \bar{I} = \frac{\beta_1(\bar{b} + \mu_1\overline{T})}{\mu_1}, \ \overline{V} = \frac{k\beta_2\bar{I}}{\mu_2},$$
$$\bar{Z} = \frac{q\beta_1\beta_3(g(\overline{T}, \bar{I}) + h(\overline{T}, \overline{V}))}{p\widetilde{\mu}}, \tag{2.1}$$

where $\overline{T}$ is given in $(\mathbf{H}_1)$, $\bar{b} = \sup_{[0,\overline{T}]} b(T)$ and $\widetilde{\mu} = \min\{\mu_1, \mu_3\}$. Let

$$\Gamma = \Big\{(\phi_1, \phi_2, \phi_3, \phi_4) \in \mathcal{C}_+^4 : \phi_1(\theta) \le \overline{T}, \phi_2(\theta) \le \bar{I}, \phi_3(\theta) \le \overline{V}, \\ \phi_4(\theta) \le \bar{Z} \text{ for all } \theta \in (-\infty, 0]\Big\}. \tag{2.2}$$

Using the same arguments as in the proof of (Li and Shu 2010, Proposition 2.1) or (Shu et al. 2013, Lemma 2.1), we can prove that system (1.1) with initial conditions in $\mathcal{C}_+^4$ is well-posed and point dissipative.

Lemma 2.1 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *are satisfied. Then the region* Γ *is positively invariant and absorbing in* $\mathcal{C}_+^4$ *for system (1.1); namely, all solutions with initial conditions in* $\mathcal{C}_+^4$ *ultimately enter* Γ.

It is easily seen that the system (1.1) possesses an infection-free equilibrium (IFE) $E_0 = (\overline{T}, 0, 0, 0)$. Following the procedure of calculating the basic reproduction number for structured models in (Thieme 2009), we linearize the system (1.1) at the IFE E_0 and derive the basic reproduction number for infection R_0 of (1.1) as

$$R_0 = R_0^v + R_0^c, \text{ where } R_0^v = \frac{k\beta_1\beta_2}{\mu_1\mu_2}\frac{\partial h(\overline{T}, 0)}{\partial V} \text{ and} \\ R_0^c = \frac{\beta_1}{\mu_1}\frac{\partial g(\overline{T}, 0)}{\partial I}. \tag{2.3}$$

Here, R_0^v denotes the number of secondly infected cells through the virus-to-cell infection and we refer it to as the basic reproduction number for virus-to-cell infection. Also, R_0^c represents the number of secondly infected cells through the cell-to-cell infection and we refer it to as the basic reproduction number for cell-to-cell infection. To investigate the existence, the uniqueness, and the stability of the equilibria, we define

$$R(T, I, V) = \frac{k\beta_1\beta_2}{\mu_1\mu_2}\frac{h(T, V)}{V} + \frac{\beta_1}{\mu_1}\frac{g(T, I)}{I} \text{ for } I, V > 0; \\ R(T, 0, 0) = \lim_{(I,V)\to(0,0)} R(T, I, V). \tag{2.4}$$

Note that $(\mathbf{H_2})$ implies that $R(\overline{T}, 0, 0) = R_0$. An immune-inactivated equilibrium (IIE) $E_1 = (T_1, I_1, V_1, 0)$ exists if $T_1, I_1, V_1 > 0$ satisfy

$$b(T_1) = g(T_1, I_1) + h(T_1, V_1) = \mu_1 I_1/\beta_1 = \mu_1\mu_2 V_1/(k\beta_1\beta_2). \tag{2.5}$$

A simple calculation gives $I_1 = \frac{\beta_1}{\mu_1}b(T_1)$ and $V_1 = \frac{k\beta_1\beta_2}{\mu_1\mu_2}b(T_1)$. Now, we introduce the function

$$G(T) = b(T) - h(T, \frac{k\beta_1\beta_2}{\mu_1\mu_2}b(T)) - g(T, \frac{\beta_1}{\mu_1}b(T)).$$

Clearly, $G(0) = b(0) > 0$, $G(\overline{T}) = 0$ and

$$G'(\overline{T}) = b'(\overline{T})\left(1 - \frac{k\beta_1\beta_2}{\mu_1\mu_2}\frac{\partial h(\overline{T},0)}{\partial V} - \frac{\beta_1}{\mu_1}\frac{\partial g(\overline{T},0)}{\partial I}\right) = b'(\overline{T})(1-R_0).$$

Note from $(\mathbf{H_1})$ that $b'(\overline{T}) < 0$. If $R_0 > 1$, we have $G'(\overline{T}) > 0$, and hence, there exists $T_1 \in (0, \overline{T})$ such that $G(T_1) = 0$. This proves the existence of an IIE E_1 when $R_0 > 1$. We claim that $R_0 > 1$ is also a necessary condition for the existence of E_1. Suppose, to the contrary, there exists an IIE $E_1 = (T_1, I_1, V_1, 0)$ while $R_0 \le 1$. Clearly, $T_1 \in (0, \overline{T})$. This, together with $(\mathbf{H_2})$, (2.3), (2.4) and (2.5), leads to

$$1 \ge R_0 = R(\overline{T}, 0, 0) \ge R(\overline{T}, I_1, V_1) > R(T_1, I_1, V_1) = 1,$$

a contradiction. Therefore, an IIE E_1 exists if and only if $R_0 > 1$. To guarantee the uniqueness of IIE, we require an additional condition. Define

$$G_b = \left\{\xi \in [0, \overline{T}] \mid (b(T) - b(\xi))(T - \xi) < 0 \text{ for } T \neq \xi,\ T \in [0, \overline{T}]\right\}. \quad (2.6)$$

Clearly, $\overline{T} \in G_b$ and $G_b \neq \emptyset$. We shall prove by contradiction that if there exists an IIE $E_1 = (T_1, I_1, V_1, 0)$ with $T_1 \in G_b$, then E_1 is the unique IIE. Suppose there exists another IIE $E_1^* = (T_1^*, I_1^*, V_1^*, 0)$. We assume without loss of generality that $T_1^* < T_1$. Then we have $b(T_1) < b(T_1^*)$. By (2.5), we further have $I_1 < I_1^*$ and $V_1 < V_1^*$. On account of $(\mathbf{H_2})$ and (2.4), we obtain

$$1 = R(T_1^*, I_1^*, V_1^*) < R(T_1, I_1^*, V_1^*) \le R(T_1, I_1, V_1) = 1,$$

a contradiction. Hence, E_1 is the unique IIE of (1.1).

An immune-activated equilibrium (IAE) $E_2 = (T_2, I_2, V_2, Z_2)$ exists if $T_2, I_2, V_2, Z_2 > 0$ satisfy the following equilibrium equations:

$$b(T_2) = g(T_2, I_2) + h(T_2, V_2) = \frac{\mu_1 I_2 + p I_2 Z_2}{\beta_1},\ I_2 = \frac{\mu_3}{q\beta_3},\ V_2 = \frac{k\beta_2\mu_3}{q\beta_3\mu_2}. \quad (2.7)$$

Define $N(T) = b(T) - g(T, I_2) - h(T, V_2)$. Clearly, $N(0) = b(0) > 0$ and $N(\overline{T}) < 0$. Thus there exists $T_2 \in (0, \overline{T})$ such that $N(T_2) = 0$. It then follows from (2.4) and (2.7) that

$$Z_2 = \frac{\mu_1}{p}\left[\frac{\beta_1(g(T_2, I_2) + h(T_2, V_2))}{\mu_1 I_2} - 1\right] = \frac{\mu_1}{p}[R(T_2, I_2, V_2) - 1] > 0 \quad (2.8)$$

if and only if $R(T_2, I_2, V_2) > 1$. Hence, an IAE E_2 exists if and only if $R_1 > 1$, where

$$R_1 := R(T_2, I_2, V_2) = \frac{k\beta_1\beta_2}{\mu_1\mu_2}\frac{h(T_2, V_2)}{V_2} + \frac{\beta_1}{\mu_1}\frac{g(T_2, I_2)}{I_2} < R(\overline{T}, 0, 0) = R_0. \quad (2.9)$$

Next, we show that $E_2 = (T_2, I_2, V_2, Z_2)$ is the unique IAE if $T_2 \in G_b$. Suppose, to the contrary, there exists another IAE $E_2^* = (T_2^*, I_2^*, V_2^*, Z_2^*)$ with $T_2^* < T_2$, $I_2 = I_2^*$ and $V_2 = V_2^*$. Then we have $b(T_2) < b(T_2^*)$. It follows from (2.7) that $Z_2 < Z_2^*$. On the other hand, since $T_2^* < T_2$, $I_2 = I_2^*$ and $V_2 = V_2^*$, we have $R(T_2^*, I_2^*, V_2^*) < R(T_2, I_2, V_2)$. This, together with (2.8), yields $Z_2^* < Z_2$, a contradiction. Therefore, E_2 is the unique IAE of (1.1) if $T_2 \in G_b$.

To summarize, we have the following results on the existence and uniqueness of equilibria for system (1.1).

Theorem 2.1 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *are satisfied. Let* R_0, R_1 *and* G_b *be defined in (2.3), (2.9) and (2.6), respectively.*

(i) If $R_0 \le 1$, *then the infection-free equilibrium (IFE)* $E_0 = (\overline{T}, 0, 0, 0)$ *is the unique equilibrium for (1.1).*

(ii) If $R_1 \le 1 < R_0$, *then, besides* E_0, *there exists at least one immune-inactivated equilibrium (IIE)* $E_1 = (T_1, I_1, V_1, 0)$ *for (1.1), and there are no immune-activated equilibria (IAE); if further* $T_1 \in G_b$, *then* E_1 *is the unique IIE.*

(iii) If $R_1 > 1$, *then, besides* E_0 *and at least one IIE* E_1, *there is at least one IAE* $E_2 = (T_2, I_2, V_2, Z_2)$ *for (1.1); if further* $T_2 \in G_b$, *then* E_2 *is the unique IAE.*

Based on the linearized equations of (1.1) at the IIE E_1, we obtain the basic reproduction number for immune response

$$R_{IM} = \frac{q\beta_3 I_1}{\mu_3}, \tag{2.10}$$

which is defined only when $R_0 > 1$. This is because the equation (2.5) has a positive root $I_1 > 0$ if and only if $R_0 > 1$.

Lemma 2.2 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *are satisfied,* $R_0 > 1$ *and* $T_1 \in G_b$, *where* G_b *is defined in (2.6). Then there exist positive* T_1, I_1, V_1 *satisfying (2.5) and positive* T_2, I_2, V_2 *satisfying (2.7). Moreover, we have*

$$Sign(T_2 - T_1) = Sign(I_1 - I_2) = Sign(V_1 - V_2) = Sign(R_1 - 1) = Sign(R_{IM} - 1).$$

Proof From (2.5) and (2.7), we have $V_1 - V_2 = k\beta_2(I_1 - I_2)/\mu_2$. Since $T_1 \in G_b$, we obtain $(b(T_2) - b(T_1))(T_2 - T_1) < 0$. This, together with the strictly monotonicity of $g(T, I)$ and $h(T, V)$, and $b(T_i) = g(T_i, I_i) + h(T_i, V_i)$ with $i = 1, 2$, yields $\mathrm{Sign}(T_2 - T_1) = \mathrm{Sign}(I_1 - I_2) = \mathrm{Sign}(V_1 - V_2)$. By $(\mathbf{H_2})$ the function $R(T, I, V)$ is strictly increasing in T and nonincreasing in I and V. This, together with $R_1 - 1 = R(T_2, I_2, V_2) - R(T_1, I_1, V_1)$ from (2.5) and (2.9), indicates that $\mathrm{Sign}(R_1 - 1) = \mathrm{Sign}(T_2 - T_1)$.

By the equilibrium equations (2.5) and (2.7), we can rewrite R_1 and R_{IM} as

$$R_1 = \frac{q\beta_1\beta_3}{\mu_1\mu_3} b(T_2) \quad \text{and} \quad R_{IM} = \frac{q\beta_1\beta_3}{\mu_1\mu_3} b(T_1).$$

This, together with $(b(T_2)-b(T_1))(T_2-T_1)<0$ and $\text{Sign}(R_1-1)=\text{Sign}(T_2-T_1)$, implies that $\text{Sign}(T_2-T_1)=\text{Sign}(R_{IM}-R_1)=\text{Sign}(R_1-1)=\text{Sign}(R_{IM}-1)$. This ends the proof. □

Remark 2.1 In view of Lemma 2.2, the two threshold parameters R_1 and R_{IM} are equivalent. Since R_{IM} is more biologically relevant, we will always use R_{IM} as the threshold parameter. In particular, the last two conditions in the statement of Theorem 2.1, namely, $R_1 \le 1 < R_0$ and $R_1 > 1$, can be restated as $R_{IM} \le 1 < R_0$ and $R_{IM} > 1$, respectively.

3 Global stability analysis

In this section, we establish the global stability of equilibria of (1.1) by analyzing the distribution of the characteristic roots, constructing suitable Lyapunov functionals, and applying uniform persistence theory (Hale and Waltman 1989).

3.1 Global stability of the IFE E_0

Before proving the global stability of E_0, we first analyze its local stability. The characteristic equation associated with the linearization of (1.1) at $E_0=(\overline{T},0,0,0)$ is

$$(\lambda+\mu_3)(\lambda-b'(\overline{T}))F_0(\lambda)=0, \tag{3.1}$$

where

$$F_0(\lambda)=(\lambda+\mu_2)\left(\lambda+\mu_1-\frac{\partial g(\overline{T},0)}{\partial I}\int_0^\infty f_1(\tau)e^{-\lambda\tau}d\tau\right)$$
$$-k\frac{\partial h(\overline{T},0)}{\partial V}\int_0^\infty f_1(\tau)e^{-\lambda\tau}d\tau\int_0^\infty f_2(\tau)e^{-\lambda\tau}d\tau.$$

Two eigenvalues are $\lambda_1=-\mu_3$, $\lambda_2=b'(\overline{T})<0$, and all other eigenvalues are determined by $F_0(\lambda)=0$, which can be rewritten as

$$1+\frac{\lambda}{\mu_1}=\frac{\partial g(\overline{T},0)}{\partial I}\frac{\int_0^\infty f_1(\tau)e^{-\lambda\tau}d\tau}{\mu_1}$$
$$+\frac{\partial h(\overline{T},0)}{\partial V}\frac{k\int_0^\infty f_1(\tau)e^{-\lambda\tau}d\tau\int_0^\infty f_2(\tau)e^{-\lambda\tau}d\tau}{\mu_1(\lambda+\mu_2)}. \tag{3.2}$$

If $R_0<1$, then $F_0(0)=\mu_1\mu_2(1-R_0)>0$, and thus 0 is not an eigenvalue. Let $\lambda=a+bi$ be an eigenvalue of (3.2). We claim that $a<0$ if $R_0<1$. Assume to the contrary that $a\ge 0$. Then it follows from (3.2) that

$$1<\left|1+\frac{\lambda}{\mu_1}\right|\le R_0^v\left|\frac{\mu_2}{\lambda+\mu_2}\right|+R_0^c<R_0^v+R_0^c=R_0<1,$$

a contradiction. Thus, we conclude that the IFE E_0 is locally asymptotically stable if $R_0 < 1$.

If $R_0 > 1$, then $F_0(0) = \mu_1\mu_2(1 - R_0) < 0$ and $\lim_{\lambda\to\infty} F_0(\lambda) = \infty$. Thus, there exists at least one positive eigenvalue, which implies that E_0 is unstable. We now have the following result.

Lemma 3.1 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *hold. The IFE* E_0 *of (1.1) is locally asymptotically stable if* $R_0 < 1$ *and unstable if* $R_0 > 1$.

For the critical case $R_0 = 1$, we have $F_0(0) = 0$ and 0 is a simple eigenvalue. Using a similar argument as above we can show that all other eigenvalues have negative real parts. Now, we examine the local stability of E_0 by using the center manifold theory and the normal forms.

Lemma 3.2 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *hold. Denote*

$$\chi := \frac{\partial^2 g(\overline{T},0)}{\partial T\partial I} + \left(\frac{k\beta_2}{\mu_2}\right)\frac{\partial^2 h(\overline{T},0)}{\partial T\partial V} - \frac{\beta_1 b'(\overline{T})}{2\mu_1}\left[\frac{\partial^2 g(\overline{T},0)}{\partial I^2} + \left(\frac{k\beta_2}{\mu_2}\right)^2\frac{\partial^2 h(\overline{T},0)}{\partial V^2}\right]. \tag{3.3}$$

If $R_0 = 1$, *then the IFE* E_0 *of (1.1) is locally asymptotically stable when* $\chi > 0$ *and unstable when* $\chi < 0$.

Proof Let $\Lambda = \{\lambda \in \mathbb{C}, \lambda \text{ is an eigenvalue of (3.1) with } \mathrm{Re}\lambda = 0\}$. Then $\Lambda = \{0\}$ if $R_0 = 1$, and (1.1) satisfies the nonresonance condition relative to Λ. Let $u(t) = (u_1(t), u_2(t), u_3(t), u_4(t))^T = (\overline{T} - T(t), I(t), V(t), Z(t))^T$. By using the standard notation in delay differential equations $u_t(\theta) = u(t+\theta)$, system (1.1) can be rewritten as an abstract ODE

$$\dot{u}(t) = Au_t + R(u_t) \tag{3.4}$$

on $\mathcal{C}^4$, where A is a linear operator defined as $(A\phi)(\theta) = \phi'(\theta)$ for $\theta \in [-\infty, 0)$ and

$$(A\phi)(0) = \begin{pmatrix} -b'(\overline{T})\phi_1(0) + \frac{\partial g(\overline{T},0)}{\partial I}\phi_2(0) + \frac{\partial h(\overline{T},0)}{\partial V}\phi_3(0) \\ -\mu_1\phi_2(0) + \frac{\partial g(\overline{T},0)}{\partial I}\int_0^\infty f_1(\tau)\phi_2(-\tau)d\tau + \frac{\partial h(\overline{T},0)}{\partial V}\int_0^\infty f_1(\tau)\phi_3(-\tau)d\tau \\ -\mu_2\phi_3(0) + k\int_0^\infty f_2(\tau)\phi_2(-\tau)d\tau \\ -\mu_3\phi_4(0) \end{pmatrix},$$

and R is a nonlinear operator defined as $(R(\phi))(\theta) = 0$ for $\theta \in [-\infty, 0)$ and

$$\begin{aligned}&(R(\phi))(0)\\ &= \begin{pmatrix} -b(\overline{T} - \phi_1(0)) + b'(\overline{T})\phi_1(0) + g(\overline{T} - \phi_1(0), \phi_2(0)) - \frac{\partial g(\overline{T},0)}{\partial I}\phi_2(0) \\ +h(\overline{T} - \phi_1(0), \phi_3(0)) - \frac{\partial h(\overline{T},0)}{\partial V}\phi_3(0) \\ \int_0^\infty f_1(\tau)g(\overline{T} - \phi_1(-\tau), \phi_2(-\tau))d\tau - \frac{\partial g(\overline{T},0)}{\partial I}\int_0^\infty f_1(\tau)\phi_2(-\tau)d\tau \\ +\int_0^\infty f_1(\tau)h(\overline{T} - \phi_1(-\tau), \phi_3(-\tau))d\tau - \frac{\partial h(\overline{T},0)}{\partial V}\int_0^\infty f_1(\tau)\phi_3(-\tau)d\tau - p\phi_2(0)\phi_4(0) \\ 0 \\ q\int_0^\infty f_3(\tau)\phi_2(-\tau)\phi_4(-\tau)d\tau \end{pmatrix}\end{aligned}$$

for $\phi = (\phi_1, \phi_2, \phi_3, \phi_4)^T \in \mathcal{C}^4$. For $\psi = (\psi_1, \psi_2, \psi_3, \psi_4) \in (C([0, \infty), \mathbb{R}))^4$ and $\phi \in \mathcal{C}^4$, we introduce a bilinear form

$$\begin{aligned}\langle \psi, \phi \rangle = \psi(0)\phi(0) + \int_0^\infty k f_2(\tau) \int_{-\tau}^0 \psi_3(\theta + \tau)\phi_2(\theta) d\theta d\tau \\ + \int_0^\infty f_1(\tau) \int_{-\tau}^0 \psi_2(\theta + \tau) \left(\frac{\partial g(\overline{T}, 0)}{\partial I} \phi_2(\theta) + \frac{\partial h(\overline{T}, 0)}{\partial V} \phi_3(\theta) \right) d\theta d\tau.\end{aligned}$$

We then choose $\varphi = (1, \varphi_2, \varphi_3, 0)^T$ and $\psi = (0, 1, \psi_3, 0)$, respectively, to be the right and left eigenvectors of the linear operator A corresponding to the eigenvalue 0, where

$$\varphi_2 = \frac{\beta_1 b'(\overline{T})}{\mu_1} < 0, \ \varphi_3 = \frac{k\beta_2}{\mu_2}\varphi_2 < 0, \ \psi_3 = \frac{\beta_1}{\mu_2} \frac{\partial h(\overline{T}, 0)}{\partial V} > 0 \text{ and } \hat{a} = \langle \psi, \varphi \rangle < 0. \tag{3.5}$$

We make the decomposition $u_t = z\varphi + y$ such that $\langle \psi, y \rangle = 0$. It is readily seen that

$$\langle \psi, \dot{u}_t \rangle = \dot{z}\langle \psi, \varphi \rangle + \langle \psi, \dot{y} \rangle = \dot{z}\langle \psi, \varphi \rangle.$$

Since $A\varphi = 0$ and $\langle \psi, Ay \rangle = 0$, we have

$$\langle \psi, \dot{u}_t \rangle = \langle \psi, Au_t \rangle + \langle \psi, R(u_t) \rangle = \langle \psi, R(u_t) \rangle = \langle \psi, R(z\varphi + y) \rangle.$$

Coupling the above two equations gives

$$\dot{z}\langle \psi, \varphi \rangle = \langle \psi, R(z\varphi + y) \rangle = \psi(R(z\varphi + y))(0) = (R(z\varphi + y))_2(0).$$

If the initial value is a small perturbation of E_0, then z is also small with a positive initial value $z(0)$ and $y = O(z^2)$. By Taylor expansion, we derive the following normal form of (3.4) at origin

$$\begin{aligned}\hat{a}\dot{z} &= \int_0^\infty f_1(\tau) \left(\frac{1}{2} \frac{\partial^2 g(\overline{T}, 0)}{\partial I^2} (z\varphi_2 + y_2(-\tau))^2 - \frac{\partial^2 g(\overline{T}, 0)}{\partial T \partial I} (z + y_1(-\tau))(z\varphi_2 + y_2(-\tau)) \right. \\ &\quad \left. + \frac{1}{2} \frac{\partial^2 h(\overline{T}, 0)}{\partial V^2} (z\varphi_3 + y_3(-\tau))^2 - \frac{\partial^2 h(\overline{T}, 0)}{\partial T \partial V} (z + y_1(-\tau))(z\varphi_3 + y_3(-\tau)) \right) d\tau + O(z^3) \\ &= \beta_1 \left(\frac{\varphi_2^2}{2} \frac{\partial^2 g(\overline{T}, 0)}{\partial I^2} - \varphi_2 \frac{\partial^2 g(\overline{T}, 0)}{\partial T \partial I} + \frac{\varphi_3^2}{2} \frac{\partial^2 h(\overline{T}, 0)}{\partial V^2} - \varphi_3 \frac{\partial^2 h(\overline{T}, 0)}{\partial T \partial V} \right) z^2 + O(z^3) \\ &= -\beta_1 \varphi_2 \chi z^2 + O(z^3),\end{aligned}$$

where χ is defined in (3.3). Note from (3.5) that $\hat{a} < 0$ and $\varphi_2 < 0$. Thus, the zero solution of the normal form equation with positive initial value is locally asymptotically stable if $\chi > 0$ and unstable if $\chi < 0$. This proves the lemma. □

Remark 3.1 If we choose the bilinear (mass-action) incidence rates $g(T, I) = c_1 TI$ and $h(T, V) = c_2 TV$ in model (1.1), then $\chi = c_1 + c_2 k\beta_2/\mu_2 > 0$ and the IFE E_0 is locally asymptotically stable when $R_0 = 1$. If we choose the standard incidence rates $g(T, I) = c_1 TI/(T + I)$ and $h(T, V) = c_2 TV/(T + V)$ in model (1.1), then $\chi = \beta_1 b'(\overline{T})[c_1 + c_2(k\beta_2/\mu_2)^2]/(\mu_1 \overline{T}) < 0$ and the IFE E_0 is unstable when $R_0 = 1$.

For simplicity, we denote

$$h_1(T) = \lim_{V\to 0^+} \frac{h(T, V)}{V} = \frac{\partial h(T, 0)}{\partial V} > 0 \text{ and } g_1(T) = \lim_{I\to 0^+} \frac{g(T, I)}{I} = \frac{\partial g(T, 0)}{\partial I} > 0 \tag{3.6}$$

for any $T > 0$, where $h_1(\overline{T})$ and $g_1(\overline{T})$ are the virus-to-cell and cell-to-cell infection rates, respectively.

Theorem 3.1 *Assume that* $(\mathbf{H_1})$-$(\mathbf{H_2})$ *are satisfied. If* $R_0 < 1$, *then the IFE* $E_0 = (\overline{T}, 0, 0, 0)$ *of (1.1) is globally asymptotically stable in* $\mathcal{C}_+^4$.

Proof In view of Lemmas 2.1 and 3.1, we only need to show that E_0 is globally attractive in Γ. Define a Lyapunov functional $L_0 : \Gamma \to \mathbb{R}$ by

$$\begin{aligned} L_0(\phi_1, \phi_2, \phi_3, \phi_4) = \phi_2(0) + \frac{\beta_1}{\mu_2} h_1(\overline{T})\phi_3(0) + \frac{k\beta_1}{\mu_2} h_1(\overline{T}) \int_0^\infty f_2(\tau) \int_{-\tau}^0 \phi_2(\theta) d\theta d\tau \\ + \int_0^\infty f_1(\tau) \int_{-\tau}^0 \Big(h(\phi_1(\theta), \phi_3(\theta)) + g(\phi_1(\theta), \phi_2(\theta))\Big) d\theta d\tau, \end{aligned}$$

where β_i $(i = 1, 2)$ and $h_1(\overline{T})$ are defined in (2.1) and (3.6), respectively. Calculating the time derivative of L_0 along the solution of (1.1), we obtain

$$\begin{aligned} L_0' = \beta_1 \big(h(T(t), V(t)) - h_1(\overline{T})V(t)\big) - pI(t)Z(t) \\ + \mu_1 I(t)\Big(\frac{k\beta_1\beta_2}{\mu_1\mu_2} h_1(\overline{T}) + \frac{\beta_1 g(T(t), I(t))}{\mu_1 I(t)} - 1\Big). \end{aligned}$$

The definitions of $h_1(T)$ and $g_1(T)$ in (3.6) and $(\mathbf{H_2})$ imply that $h(T(t), V(t)) \le h_1(\overline{T})V(t)$ and $g(T(t), I(t)) \le g_1(\overline{T})I(t)$ for $0 \le T \le \overline{T}$ and $I, V \ge 0$. This, together with the definition of R_0 in (2.3), yields

$$L_0' \le \beta_1 \big(h(T(t), V(t)) - h_1(\overline{T})V(t)\big) - pI(t)Z(t) + \mu_1 I(t)(R_0 - 1) \le 0.$$

Furthermore, $L_0' = 0$ if and only if $(T(t), I(t), V(t), Z(t)) \equiv (\overline{T}, 0, 0, 0)$. Thus, the maximal compact invariant set in $\{L_0' = 0\}$ is the singleton $\{E_0\}$. By the LaSalle invariance principle (Hale and Verduyn Lunel 1993, Theorem 5.3.1), E_0 is globally attractive in Γ. Since Γ is absorbing in $\mathcal{C}_+^4$, we conclude that E_0 is globally attractive in $\mathcal{C}_+^4$. It then follows from Lemma 3.1 that the IFE E_0 is globally asymptotically stable in $\mathcal{C}_+^4$. □

Corollary 3.1 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *hold and that* $R_0 = 1$. *Then the IFE* E_0 *is globally asymptotically stable in* $\mathcal{C}_+^4$ *provided that* $\chi > 0$.

3.2 Global stability of the IIE E_1

In this subsection, we investigate the stability of the IIE E_1, which exists only when $R_0 > 1$ and is unique if $T_1 \in G_b$. Denote $c_0 = \partial h(T_1, V_1)/\partial T + \partial g(T_1, I_1)/\partial T$, $c_1 = \partial g(T_1, I_1)/\partial I$ and $c_2 = \partial h(T_1, V_1)/\partial V$. The characteristic equation of the linearized system of (1.1) at $E_1 = (T_1, I_1, V_1, 0)$ is

$$F_1(\lambda)F_2(\lambda) = 0, \tag{3.7}$$

where $F_1(\lambda) = \lambda + \mu_3 - qI_1 \int_0^\infty f_3(\tau)e^{-\lambda\tau}d\tau$ and

$$\begin{aligned} F_2(\lambda) = &(\lambda - b'(T_1) + c_0)(\lambda + \mu_1)(\lambda + \mu_2) \\ &- (\lambda - b'(T_1))(c_1(\lambda + \mu_2) + kc_2 \int_0^\infty f_2(\tau)e^{-\lambda\tau}d\tau) \int_0^\infty f_1(\tau)e^{-\lambda\tau}d\tau. \end{aligned}$$

From the proof of (Shu et al. 2013, Theorem 3.2), we have the following three results: (1) all roots of $F_1(\lambda) = 0$ have negative real parts if and only if $R_{IM} < 1$, i.e., $0 < q\beta_3 I_1 < \mu_3 = q\beta_3 I_2$; (2) there exists at least one positive root if $R_{IM} > 1$; and (3) 0 is a simple roots and all other roots have negative real parts if $R_{IM} = 1$.

We now claim that all roots of $F_2(\lambda) = 0$ have negative real parts. Otherwise, suppose that $\lambda = a + bi$ is a zero of $F_2(\lambda)$ satisfying $a \geq 0$. Note that $F_2(\lambda) = 0$ can be rewritten as $P_L(\lambda) = P_R(\lambda)$, where

$$\begin{aligned} P_L(\lambda) &= \frac{\lambda - b'(T_1) + c_0}{\lambda - b'(T_1)}(\frac{\lambda}{\mu_1} + 1), \\ P_R(\lambda) &= \left(\frac{c_1}{\mu_1} + \frac{kc_2}{\mu_1(\lambda + \mu_2)} \int_0^\infty f_2(\tau)e^{-\lambda\tau}d\tau\right) \int_0^\infty f_1(\tau)e^{-\lambda\tau}d\tau. \end{aligned}$$

Clearly, $|P_L| > 1$. On the other hand, since $(\mathbf{H_2})$ implies that $0 < c_1 \leq g(T_1, I_1)/I_1$ and $0 < c_2 \leq h(T_1, V_1)/V_1$, we have

$$|P_R| \leq \frac{\beta_1}{\mu_1}\frac{g(T_1, I_1)}{I_1} + \frac{k\beta_1\beta_2}{\mu_1\mu_2}\frac{h(T_1, V_1)}{V_1} = R(T_1, I_1, V_1) = 1.$$

This leads to a contradiction. Hence, we obtain the following three results: (1) all eigenvalues of (3.7) have negative real parts if and only if $R_{IM} < 1$; (2) there exists at least one positive eigenvalue if $R_{IM} > 1$; and (3) 0 is a simple eigenvalue and all other eigenvalues have negative real parts if $R_{IM} = 1$. We use a similar argument as in the proof of Lemma 3.2 to calculate the normal form of (1.1) at E_1 when $R_{IM} = 1$. The resulting equation is given by

$$\begin{aligned} \dot{z} &= \frac{q\beta_3\widetilde{a}}{1 + qI_1 \int_0^\infty \tau f_3(\tau)d\tau}z^2 + O(z^3), \text{ where} \\ \widetilde{a} &= \frac{b'(T_1) - \partial g(T_1, I_1)/\partial T - \partial h(T_1, V_1)/\partial T}{\partial g(T_1, I_1)/\partial I + \frac{k\beta_2}{\mu_2}\partial h(T_1, V_1)/\partial V} < 0. \end{aligned}$$

Therefore, E_1 of (1.1) is locally asymptotically stable if $R_{IM} = 1$. To summarize, we have the following conclusion.

Lemma 3.3 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *hold. Let* $T_1 \in G_b$, *where* G_b *is defined in (2.6). Then the IIE* $E_1 = (T_1, I_1, V_1, 0)$ *of (1.1) is locally asymptotically stable if* $R_{IM} \leq 1 < R_0$ *and unstable if* $R_{IM} > 1$.

If $R_0 > 1$, then the infection is persistent in the sense that the virus cannot be cleared. Denote

$$X_1 = \left\{(\phi_1, \phi_2, \phi_3, \phi_4) \in \mathcal{C}_+^4 : \text{ either } I_0(\theta) > 0 \text{ or } V_0(\theta) > 0 \text{ for some } \theta \in (-\infty, 0]\right\}. \tag{3.8}$$

By using (Hale and Waltman 1989, Theorem 4.1) and (Shu et al. 2018, Lemma 4.2), we obtain the following persistence result for system (1.1).

Theorem 3.2 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *hold. If* $R_0 > 1$, *then there exists an* $\eta_0 > 0$ *such that* $\liminf_{t\to\infty} T(t) \geq \eta_0$, $\liminf_{t\to\infty} I(t) \geq \eta_0$ *and* $\liminf_{t\to\infty} V(t) \geq \eta_0$ *for any solution of (1.1) with initial condition in* X_1.

In many models, $h(T, V)$ and $g(T, I)$ are assumed to take the following forms.

$(\mathbf{H_3})$ There exist $p \in C^1(\mathbb{R}_+), h_0, g_0 \in C^2(\mathbb{R}_+)$ such that $h(T, V) = p(T)h_0(V)$ and $g(T, I) = p(T)g_0(I)$.

Theorem 3.3 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_3})$ *hold and* $T_1 \in G_b$. *Let* G_b *and* X_1 *be as defined in (2.6) and (3.8), respectively. If* $R_{IM} \leq 1 < R_0$, *then the IIE* $E_1 = (T_1, I_1, V_1, 0)$ *of (1.1) is globally asymptotically stable in* X_1.

Proof Theorem 2.1(ii) and Lemma 3.3 imply that the IIE E_1 exists uniquely and is locally asymptotically stable. We only need to show that E_1 is globally attractive in the positively invariant set X_1. Denote $u(\theta) = \theta - 1 - \ln\theta$. It is obvious that $u(\theta) \geq 0$ for $\theta > 0$ and $u(\theta) = 0$ if and only if $\theta = 1$. Define a Lyapunov functional $L_1 : X_1 \to \mathbb{R}$ as

$$\begin{aligned} L_1(\phi) = {} & \beta_1 \int_{T_1}^{\phi_1(0)} \left(1 - \frac{h(T_1, V_1)}{h(s, V_1)}\right) ds + I_1 u\left(\frac{\phi_2(0)}{I_1}\right) + \frac{\beta_1 h(T_1, V_1)}{\mu_2} u\left(\frac{\phi_3(0)}{V_1}\right) \\ & + \frac{p}{q\beta_3}\phi_4(0) + h(T_1, V_1)W_1 \\ & + g(T_1, I_1)W_2 + \frac{\beta_1 h(T_1, V_1)}{\beta_2} \int_0^\infty f_2(\tau) \int_{-\tau}^0 u\left(\frac{\phi_2(\theta)}{I_1}\right) d\theta d\tau \\ & + \frac{p}{\beta_3} \int_0^\infty f_3(\tau) \int_{-\tau}^0 \phi_2(\theta)\phi_4(\theta) d\theta d\tau, \end{aligned}$$

where $\phi = (\phi_1, \phi_2, \phi_3, \phi_4) \in X_1$ and

$$W_1 = \int_0^\infty f_1(\tau) \int_{-\tau}^0 u\left(\frac{h(\phi_1(\theta), \phi_3(\theta))}{h(T_1, V_1)}\right) d\theta d\tau,$$

$$W_2 = \int_0^\infty f_1(\tau) \int_{-\tau}^0 u\left(\frac{g(\phi_1(\theta), \phi_2(\theta))}{g(T_1, I_1)}\right) d\theta d\tau.$$

Lemma 2.1 and Theorem 3.2 imply that L_1 is well-defined in X_1. By using $(\mathbf{H}_3)$, (2.5) and (2.7), we calculate and simplify the time derivative of L_1 along solutions of (1.1) as

$$\begin{aligned}
L_1' = {} & \beta_1(b(T(t)) - b(T_1))\left(1 - \frac{p(T_1)}{p(T(t))}\right) - \beta_1 h(T_1, V_1) \\
& \left[u\left(\frac{p(T_1)}{p(T(t))}\right) + u\left(\frac{h_0(V_1)V(t)}{h_0(V(t))V_1}\right)\right] \\
& + p(I_1 - I_2)Z(t) - \beta_1 g(T_1, I_1)\left[u\left(\frac{p(T_1)}{p(T(t))}\right) + u\left(\frac{g_0(I_1)I(t)}{g_0(I(t))I_1}\right)\right] \\
& - g(T_1, I_1)\int_0^\infty f_1(\tau) u\left(\frac{g(T(t-\tau), I(t-\tau))I_1}{g(T_1, I_1)I(t)}\right) d\tau - \frac{\beta_1 h(T_1, V_1)}{\beta_2} \\
& \int_0^\infty f_2(\tau) u\left(\frac{I(t-\tau)V_1}{I_1 V(t)}\right) d\tau \\
& - h(T_1, V_1)\int_0^\infty f_1(\tau) u\left(\frac{h(T(t-\tau), V(t-\tau))I_1}{h(T_1, V_1)I(t)}\right) d\tau \\
& + \beta_1 h(T_1, V_1) W_3 + \beta_1 g(T_1, I_1) W_4,
\end{aligned}$$

where the functions p, h_0 and g_0 are defined in $(\mathbf{H}_3)$, and

$$\begin{aligned}
W_3 &= \left(\frac{h_0(V(t))}{h_0(V_1)} - 1\right)\left(1 - \frac{h_0(V_1)/V_1}{h_0(V(t))/V(t)}\right), \\
W_4 &= \left(\frac{g_0(I(t))}{g_0(I_1)} - 1\right)\left(1 - \frac{g_0(I_1)/I_1}{g_0(I(t))/I(t)}\right).
\end{aligned}$$

Assumptions $(\mathbf{H_2})$–$(\mathbf{H_3})$ imply that $p(s)$, $h_0(s)$ and $g_0(s)$ are strictly increasing, while $h_0(s)/s$ and $g_0(s)/s$ are nonincreasing for $s \geq 0$. Thus, we have $W_3 \leq 0$ and $W_4 \leq 0$ for any $I, V > 0$. Since $T_1 \in G_b$, we obtain $(b(T) - b(T_1))(1 - p(T_1)/p(T)) \leq 0$ for any $T > 0$. Furthermore, by Lemma 2.2 we have $I_1 - I_2 \leq 0$. It then follows from the positive definiteness of $u(\theta)$ that $L_1' \leq 0$ in X_1. Note that $L_1' = 0$ if and only if $(T(t), I(t), V(t), Z(t)) \equiv (T_1, I_1, V_1, 0)$. Thus, the largest compact invariant set of $\{(T_t, I_t, V_t, Z_t) \in X_1 : L_1' = 0\}$ is the singleton $\{E_1\}$. Finally, by using the LaSalle invariance principle (Hale and Verduyn Lunel 1993) and Lemma 3.3, we obtain that E_1 is globally asymptotically stable in X_1. □

3.3 Impact of the distributed delay in immune response and global stability of the IAE E_2

If $R_{IM} > 1$, we obtain from Remark 2.1, Lemmas 3.1 and 3.3 that both E_0 and E_1 are unstable and the IAE E_2 exists. Moreover, $T_2 \in G_b$ ensures the uniqueness of

the positive equilibrium E_2. It has been shown in (Shu et al. 2013; Wang et al. 2007; Yang et al. 2017) that the discrete time lag in immune response may induce sustained oscillations through Hopf bifurcation. To investigate the impact of the distributed delays in immune response on viral infections, we set $f_3(\tau) = e^{-s_3\tau}\widetilde{f}_3(\tau)$ and choose $\widetilde{f}_3(\tau)$ from the following three different types of kernels:

$$\delta(\tau - \tau_3) \text{ (delta kernel)}; \ \gamma e^{-\gamma\tau} \text{ (weak kernel)}; \ \gamma^2\tau e^{-\gamma\tau} \text{ (strong kernel)}. \quad (3.9)$$

Here, $s_3 > 0$, $\tau_3 > 0$, and $\gamma > 0$. The factor $e^{-s_3\tau}$ denotes the survival probability of the immune cell before being active. The average delay is defined as $\tau_3 = \int_0^\infty \tau\widetilde{f}_3(\tau)d\tau$. A simple calculation gives $\tau_3 = 1/\gamma$ for the weak kernel and $\tau_3 = 2/\gamma$ for the strong kernel. We choose

$$\begin{aligned} &b(T) = \alpha - dT + rT\left(1 - \frac{T}{T_m}\right), \ g(T, I) = k_1 TI, \ h(T, V) = k_2 TV, \\ &f_1(\tau) = f_2(\tau) = \delta(\tau), \end{aligned} \quad (3.10)$$

where α, d, r, T_m, k_1 and k_2 are positive constants. It is easy to verify that $(\mathbf{H_1})$–$(\mathbf{H_3})$ are satisfied. For each $i = 1, 2$, we can further show that $T_i \in G_b$ if and only if $0 \le r < \frac{d}{1-T_i/T_m}$. Note that the intrinsic growth function of target cells $b(T)$ is a monotone function for $T > 0$ if and only if $r \in [0, d]$. Now, we fix $r \in [0, \frac{d}{1-T_2/T_m})$. If $R_{IM} > 1$, then the model (1.1) admits a unique IFE E_0, a unique IIE E_1 and a unique IAE E_2. When the kernel function $\widetilde{f}_3(\tau)$ is chosen to be the delta kernel, the weak kernel, or the strong kernel, our simulation shows that a stable periodic solution exists; see Fig. 1. This indicates that the immune recruitment delay may lead to sustained oscillations.

Next, we assume $f_3(\tau) = \delta(\tau)$ and rewrite the model (1.1) as

$$\begin{aligned} T'(t) &= b(T(t)) - g(T(t), I(t)) - h(T(t), V(t)), \\ I'(t) &= \int_0^\infty f_1(\tau)(g(T(t-\tau), I(t-\tau)) + h(T(t-\tau), V(t-\tau)))d\tau \\ &\quad - \mu_1 I(t) - pI(t)Z(t), \\ V'(t) &= k\int_0^\infty f_2(\tau)I(t-\tau)d\tau - \mu_2 V(t), \\ Z'(t) &= qI(t)Z(t) - \mu_3 Z(t). \end{aligned} \quad (3.11)$$

It is easy to verify that the IAE $E_2 = (T_2, I_2, V_2, Z_2)$ of system (3.11) satisfies (2.7) with $\beta_3 = 1$.

Theorem 3.4 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *hold. Assume further that* $T_2 \in G_b$, *where* G_b *is defined in (2.6). If* $R_{IM} > 1$, *then the unique IAE* $E_2 = (T_2, I_2, V_2, Z_2)$ *of (3.11) is locally asymptotically stable.*

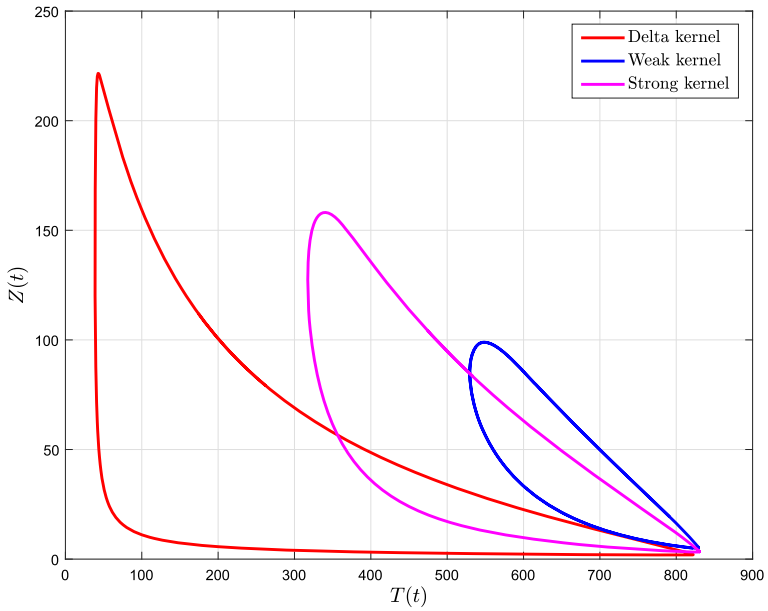

Fig. 1 Stable periodic solutions of system (1.1) with the kernel function $\widetilde{f_3}(\tau)$ being $\delta(\tau-1)$ (red), $e^{-\tau}$ (blue) and $4\tau e^{-2\tau}$ (magenta), respectively. Here, the parameter values are chosen as $\alpha = 10$, $d = 0.1$, $r = 0.2$, $T_m = 1500$, $k_1 = 0.0005$, $k_2 = 0.0007$, $\mu_1 = 0.5$, $p = 0.42$, $k = 60$, $\mu_2 = 2.6$, $q = 0.2$, $s_3 = 0.3$, and $\mu_3 = 0.1$ (color figure online)

Proof From Theorem 2.1 and $T_2 \in G_b$ we obtain the uniqueness of the positive equilibrium. Denote

$$a_0 = \partial h(T_2, V_2)/\partial T + \partial g(T_2, I_2)/\partial T > 0,\ a_1 = \partial g(T_2, I_2)/\partial I > 0,$$
$$a_2 = \partial h(T_2, V_2)/\partial V > 0.$$

The characteristic equation of the linearized system of (3.11) at the IAE E_2 is

$$F_3(\lambda) = (\lambda - b'(T_2) + a_0)(\lambda+\mu_2)\,[\lambda(\lambda+\mu_1) + pZ_2(\lambda+\mu_3)]$$
$$-\lambda(\lambda - b'(T_2))\left[a_1(\lambda+\mu_2)+a_2k\int_0^\infty f_2(\tau)e^{-\lambda\tau}d\tau\right]\int_0^\infty f_1(\tau)e^{-\lambda\tau}d\tau = 0,$$

which can be rewritten as $Q_L(\lambda) = Q_R(\lambda)$, where

$$Q_L(\lambda) = \frac{\lambda - b'(T_2) + a_0}{\lambda - b'(T_2)}\left[\frac{\lambda}{\mu_1} + 1 + \frac{pZ_2}{\mu_1}(1 + \frac{\mu_3}{\lambda})\right],$$
$$Q_R(\lambda) = \left[\frac{a_1}{\mu_1} + \frac{a_2k}{\mu_1(\lambda+\mu_2)}\int_0^\infty f_2(\tau)e^{-\lambda\tau}d\tau\right]\int_0^\infty f_1(\tau)e^{-\lambda\tau}d\tau. \qquad (3.12)$$

Since $T_2 \in G_b$, we obtain $b'(T_2) < 0$. Consequently, $F_3(0) = p\mu_2\mu_3 Z_2(a_0 - b'(T_2)) > 0$, which implies that 0 is not an eigenvalue. We now claim that all roots of

$F_3(\lambda) = 0$ have negative real parts. Otherwise, suppose that $\lambda = a + bi$ is a zero of $F_3(\lambda)$ satisfying $a \geq 0$. From the equilibrium equations (2.7) and (3.12), we have

$$|Q_L| > 1 + \frac{pZ_2}{\mu_1} = \frac{\beta_1(g(T_2, I_2) + h(T_2, V_2))}{\mu_1 I_2}.$$

It then follows from $(\mathbf{H_2})$ that $a_1 = \partial g(T_2, I_2)/\partial I \leq g(T_2, I_2)/I_2$ and $a_2 = \partial h(T_2, V_2)/\partial V \leq h(T_2, V_2)/V_2$. This, together with (3.12) and $k\beta_2 I_2 = \mu_2 V_2$, implies

$$|Q_R| \leq \frac{\beta_1}{\mu_1}\frac{g(T_2, I_2)}{I_2} + \frac{k\beta_1\beta_2}{\mu_1\mu_2}\frac{h(T_2, V_2)}{V_2} = \frac{\beta_1(g(T_2, I_2) + h(T_2, V_2))}{\mu_1 I_2},$$

a contradiction. Therefore, the unique IAE E_2 is locally asymptotically stable. □

To establish the persistence result for system (3.11), we define

$$X_2 = \Big\{(\phi_1, \phi_2, \phi_3, \phi_4) \in \mathcal{C}_+^3 \times (0, \infty) : \text{either } I_0(\theta) > 0 \text{ or } V_0(\theta) > 0 \text{ for some } \theta \in (-\infty, 0]\Big\}. \tag{3.13}$$

Theorem 3.5 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_2})$ *hold. If* $R_{IM} > 1$, *then there exists a constant* $\eta > 0$ *such that* $\liminf_{t\to\infty} T(t) \geq \eta$, $\liminf_{t\to\infty} I(t) \geq \eta$, $\liminf_{t\to\infty} V(t) \geq \eta$ *and* $\liminf_{t\to\infty} Z(t) \geq \eta$ *for any solution of (3.11) with initial condition in* X_2.

Proof From Lemma 2.2 and $R_{IM} > 1$, we have $R_0 > R_1 > 1$. It follows from Theorem 3.2 and $X_2 \subset X_1$ that $T(t)$, $I(t)$ and $V(t)$ are uniformly persistent. We only need to prove that $\liminf_{t\to\infty} Z(t) \geq \eta$ for some $\eta > 0$. According to (Hale and Waltman 1989, Theorem 4.1), it suffices to show that $W^s(E_1) \cap X_2 = \emptyset$, where $W^s(E_1)$ is the stable manifold of E_1. Suppose, to the contrary, there exists a solution $(T(t), I(t), V(t), Z(t)) \in X_2$ such that $\lim_{t\to\infty}(T(t), I(t), V(t), Z(t)) = (T_1, I_1, V_1, 0)$. Since $R_{IM} = qI_1/\mu_3 > 1$, there exist $\epsilon_1 > 0$ and $t_1 > 0$ such that $qI(t) - \mu_3 \geq \epsilon_1$ for all $t \geq t_1$. Then we have

$$Z'(t) = qI(t)Z(t) - \mu_3 Z(t) \geq \epsilon_1 Z(t) \text{ for all } t \geq t_1,$$

which contradicts to the assumption that $\lim_{t\to\infty} Z(t) = 0$. Hence, $W^s(E_1) \cap X_2 = \emptyset$. This ends the proof. □

Theorem 3.6 *Assume that* $(\mathbf{H_1})$–$(\mathbf{H_3})$ *hold and* $T_2 \in G_b$, *where* G_b *is defined in (2.6). If* $R_{IM} > 1$, *then there exists a unique IAE* $E_2 = (T_2, I_2, V_2, Z_2)$ *of (3.11) which is globally asymptotically stable in* X_2.

Proof The existence, uniqueness and locally asymptotic stability of the IAE E_2 of system (3.11) follow from Theorem 2.1(iii) and Theorem 3.4. To prove the global

attractivity of E_2 in X_2, we define a Lyapunov functional $L_2 : X_2 \to \mathbb{R}$ as

$$
\begin{aligned}
L_2(\phi) = {} & \beta_1 \int_{T_2}^{\phi_1(0)} \left(1 - \frac{h(T_2, V_2)}{h(s, V_2)}\right) ds + I_2 u\left(\frac{\phi_2(0)}{I_2}\right) + \frac{\beta_1 h(T_2, V_2)}{\mu_2} u\left(\frac{\phi_3(0)}{V_2}\right) \\
& + \frac{pZ_2}{q} u\left(\frac{Z(t)}{Z_2}\right) + h(T_2, V_2) \int_0^\infty f_1(\tau) \int_{-\tau}^0 u\left(\frac{h(\phi_1(\theta), \phi_3(\theta))}{h(T_2, V_2)}\right) d\theta d\tau \\
& + g(T_2, I_2) \int_0^\infty f_1(\tau) \int_{-\tau}^0 u\left(\frac{g(\phi_1(\theta), \phi_2(\theta))}{g(T_2, I_2)}\right) d\theta d\tau \\
& + \frac{\beta_1 h(T_2, V_2)}{\beta_2} \int_0^\infty f_2(\tau) \int_{-\tau}^0 u\left(\frac{\phi_2(\theta)}{I_2}\right) d\theta d\tau,
\end{aligned}
$$

where $\phi = (\phi_1, \phi_2, \phi_3, \phi_4) \in X_2$ and $u(\theta) = \theta - 1 - \ln\theta$. By Theorem 3.5, L_2 is well-defined in X_2. On account of $(\mathbf{H}_3)$ and the equilibrium equations (2.7), we calculate and simplify the time derivative of L_2 along solutions of (3.11) as

$$
\begin{aligned}
L_2' = {} & \beta_1 (b(T(t)) - b(T_2)) \left(1 - \frac{p(T_2)}{p(T(t))}\right) \\
& - \beta_1 h(T_2, V_2) \left[u\left(\frac{p(T_2)}{p(T(t))}\right) + u\left(\frac{h_0(V_2) V(t)}{h_0(V(t)) V_2}\right)\right] \\
& - \beta_1 g(T_2, I_2) \left[u\left(\frac{p(T_2)}{p(T(t))}\right) + u\left(\frac{g_0(I_2) I(t)}{g_0(I(t)) I_2}\right)\right] \\
& - \frac{\beta_1 h(T_2, V_2)}{\beta_2} \int_0^\infty f_2(\tau) u\left(\frac{I(t-\tau) V_2}{I_2 V(t)}\right) d\tau \\
& - g(T_2, I_2) \int_0^\infty f_1(\tau) u\left(\frac{g(T(t-\tau), I(t-\tau)) I_2}{g(T_2, I_2) I(t)}\right) d\tau \\
& - h(T_2, V_2) \int_0^\infty f_1(\tau) u\left(\frac{h(T(t-\tau), V(t-\tau)) I_2}{h(T_2, V_2) I(t)}\right) d\tau \\
& + \beta_1 h(T_2, V_2) W_5 + \beta_1 g(T_2, I_2) W_6,
\end{aligned}
$$

where p, h_0, g_0 are given in $(\mathbf{H}_3)$, and

$$
\begin{aligned}
W_5 &= \left(\frac{h_0(V(t))}{h_0(V_2)} - 1\right) \left(1 - \frac{h_0(V_2)/V_2}{h_0(V(t))/V(t)}\right), \\
W_6 &= \left(\frac{g_0(I(t))}{g_0(I_2)} - 1\right) \left(1 - \frac{g_0(I_2)/I_2}{g_0(I(t))/I(t)}\right).
\end{aligned}
$$

Similar as in the proof of Theorem 3.3, we obtain from $(\mathbf{H}_3)$ and $T_2 \in G_b$ that $L_2' \le 0$. Moreover, the largest compact invariant set of $\{(T_t, I_t, V_t, Z(t)) \in X_2 : L_2' = 0\}$ is the singleton $\{E_2\}$. Therefore, the global stability of the IAE E_2 in X_2 follows from the LaSalle invariance principle (Hale and Verduyn Lunel 1993). □

4 Viral dynamics beyond stability: Numerical explorations

In this section, we conduct numerical simulations to explore the viral dynamics beyond stability. As an illustrative example, we choose bilinear incidence rates and quadratic growth rate in (1.1) and consider the following system.

$$\begin{aligned}
T'(t) &= \lambda - dT(t) + rT(t)\left(1 - \frac{T(t)}{T_m}\right) - \alpha_1 T(t)V(t) - \alpha_2 T(t)I(t),\\
I'(t) &= \alpha_1 \int_0^\infty e^{-s_1\tau} \widetilde{f}_1(\tau) T(t-\tau)V(t-\tau)d\tau\\
&\quad + \alpha_2 \int_0^\infty e^{-s_1\tau} \widetilde{f}_1(\tau) T(t-\tau)I(t-\tau)d\tau - \mu_1 I(t) - pI(t)Z(t),\\
V'(t) &= k\int_0^\infty e^{-s_2\tau} \widetilde{f}_2(\tau) I(t-\tau)d\tau - \mu_2 V(t),\\
Z'(t) &= q\int_0^\infty e^{-s_3\tau} \widetilde{f}_3(\tau) I(t-\tau)Z(t-\tau)d\tau - \mu_3 Z(t).
\end{aligned} \tag{4.1}$$

For each $i = 1, 2, 3$, we choose the density function $\widetilde{f}_i(\tau)$ from the following three types of kernels:

$$\delta(\tau - \tau_i) \text{ (delta kernel)};\ \gamma_i e^{-\gamma_i \tau} \text{ (weak kernel)};\ \gamma_i^2 \tau e^{-\gamma_i \tau} \text{ (strong kernel)}.$$

The parameter values are taken as $\lambda = 10, d = 0.1, r = 0.2, T_m = 1500, \alpha_1 = 0.0005$, $\alpha_2 = 0.0007, s_1 = 0.4, \mu_1 = 0.5, p = 0.42, k = 60, s_2 = 0.28, \mu_2 = 2.6, q = 0.2$, $s_3 = 0.3, \mu_3 = 0.1$ unless otherwise specified.

4.1 Effects of delay in CTL immune cell recruitment on viral dynamics

If there is no delay in the recruitment of CTL immune cell, i.e., $\widetilde{f}_3(\tau) = \delta(\tau)$ in (4.1), then we obtain from Theorem 3.6 that the IAE E_2 is globally asymptotically stable. If $\widetilde{f}_3(\tau)$ is a discrete delta kernel with a positive time delay or a gamma distributed kernel, then we have observed from Fig. 1 that there may exist a sustained periodic solution. Now, we further explore the rich viral dynamics by choosing the average delay $\tau_3 = \int_0^\infty \tau \widetilde{f}_3(\tau)d\tau$ as the bifurcation parameter. We plot the bifurcation diagrams in Fig. 2, where the three kernels $\widetilde{f}_i(\tau)$ with $i = 1, 2, 3$ are chosen as the delta kernels in Fig. 2(a), the weak kernels in Fig. 2(b), and the strong kernels in Fig. 2(c). For the weak and strong kernel cases, we observe stability switch of the IAE E_2 and a stable periodic solution exists when E_2 is unstable. Similar patterns of onset and termination of Hopf bifurcations were observed for other biological models in the literature (Liu et al. 2015; Shu et al. 2020). For the delta kernel case, chaotic behavior is observed before the IAE E_2 regains its stability. In the simulations, we have chosen a large logistic growth rate $r = 3$. If $r = 0.2$, then the bifurcation diagrams for all three

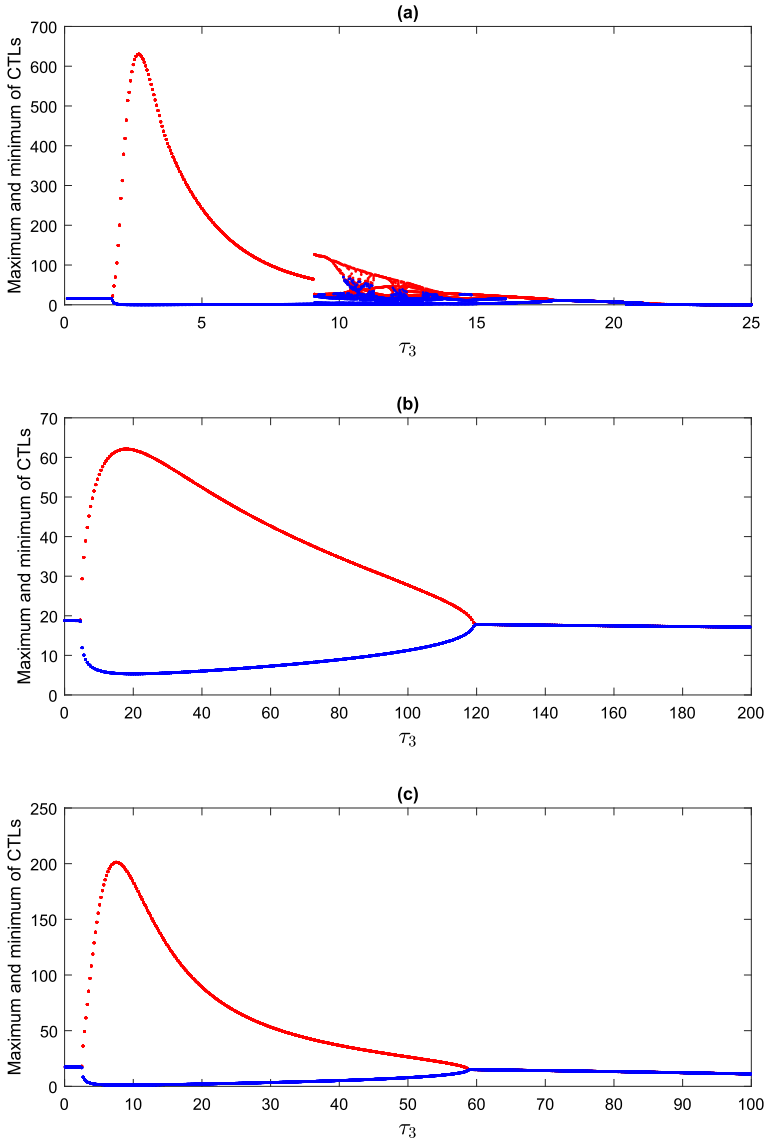

Fig. 2 Bifurcation diagrams of (4.1) with τ_3 as the bifurcation parameter. Three types of kernels are chosen: **a** delta kernels: $\widetilde{f}_i(\tau) = \delta(\tau - \tau_i)$ with $i = 1, 2, 3$, where $\tau_1 = 1$, $\tau_2 = 2$, and τ_3 varies; **b** weak kernels: $\widetilde{f}_i(\tau) = \gamma_i e^{-\gamma_i \tau}$ with $i = 1, 2, 3$, where $\gamma_1 = 1$ and $\gamma_2 = 0.5$, and $\tau_3 = 1/\gamma_3$ varies; **c** strong kernels: $f_i(\tau) = \gamma_i^2 \tau e^{-\gamma_i \tau}$ with $i = 1, 2, 3$, where $\gamma_1 = 2$ and $\gamma_2 = 1$, and $\tau_3 = 2/\gamma_3$ varies

cases are similar and we observe no chaotic behavior for the delta kernel case. Our simulation implies that the weak or strong kernel may not induce chaos but the delta kernel (with a discrete delay) may induce chaos if the logistic growth rate r is large.

4.2 Contributions of virus-to-cell and cell-to-cell modes to the production of infected cells

In this subsection, we examine contributions from both virus-to-cell and cell-to-cell modes to the production of newly infected cells. For each $i = 1, 2, 3$, the kernel $\widetilde{f}_i(\tau)$ has three choices. So, there are a total of 27 combinations. As an illustration, we choose only weak kernels: $\widetilde{f}_i(\tau) = \gamma_i e^{-\gamma_i \tau}$ for $i = 1, 2, 3$. We further fix $\gamma_1 = 1$ and $\gamma_3 = 0.2$, and choose four different values of γ_2: 1/20, 1/40, 1/60, and 1/80, respectively. The contributions of virus-to-cell and cell-to-cell infections are calculated by the two ratios $\frac{I_1(t)}{I_1(t)+I_2(t)}$ and $\frac{I_2(t)}{I_1(t)+I_2(t)}$, respectively, where $I_1(t)$ and $I_2(t)$ are determined by

$$\begin{aligned} I_1'(t) &= \alpha_1 \int_0^\infty e^{-(s_1+1)\tau} T(t-\tau) V(t-\tau) d\tau, \\ I_2'(t) &= \alpha_2 \int_0^\infty e^{-(s_1+1)\tau} T(t-\tau) I(t-\tau) d\tau, \end{aligned} \tag{4.2}$$

where $\alpha_1 = 0.0005$, $\alpha_2 = 0.0007$, and $s_1 = 0.4$ are fixed. The dynamics of the percentages of infections from two modes are plotted in Fig. 3, where four different values of γ_2 are chosen. The average delay of virion maturation $\tau_2 = 1/\gamma_2$ takes the values 20, 40, 60, and 80, respectively. We observe from Fig. 3 that the percentage of infection from cell-to-cell mode is increasing as τ_2 increases. This is biologically reasonable because a larger virion maturation delay reduces the survival probability of the virus and hence decreases the percentage of infection from virus-to-cell mode. One can also derive from (2.1) and (2.3) that both the survival probability during virion maturation $\beta_2 = \gamma_2/(s_2 + \gamma_2) = 1/(1 + s_2\tau_2)$ and the basic reproduction number for virus-to-cell infection R_0^v are decreasing in τ_2.

We also used different types of kernels for the cell-to-cell and virus-to-cell modes to address the differences in the mechanisms of these two modes. We observed a similar phenomenon that the percentage of infection from cell-to-cell mode will increase and the percentage of infection from virus-to-cell mode will decrease if we increase the average delay of virion maturation (figures not shown here).

Two other key parameters that affect the contributions of the two modes are α_1 and α_2. As indicated in Fig. 4, α_1 has a positive effect on the contribution from the virus-to-cell mode, while α_2 positively affects the contribution of the cell-to-cell mode.

4.3 Sensitive analysis

In this subsection, we employ the partial rank correlation coefficients (PRCCs) method to perform a sensitivity analysis of the basic reproduction numbers R_0 and R_{IM} for (4.1). The sign of PRCCs indicates the positive or negative correlation between the basic reproduction numbers and model parameters (Guo et al. 2020). The value of PRCCs measures the strength of the correlation; namely, a strong correlation occurs if |PRCC| >0.4, a moderate correlation happens if $0.2<$ |PRCC| <0.4, and the correlation is weak if $0<$ |PRCC| <0.2. We observe that the parameters T_m, α_1, and k have strong and positive impact on the basic reproduction number of infection R_0;

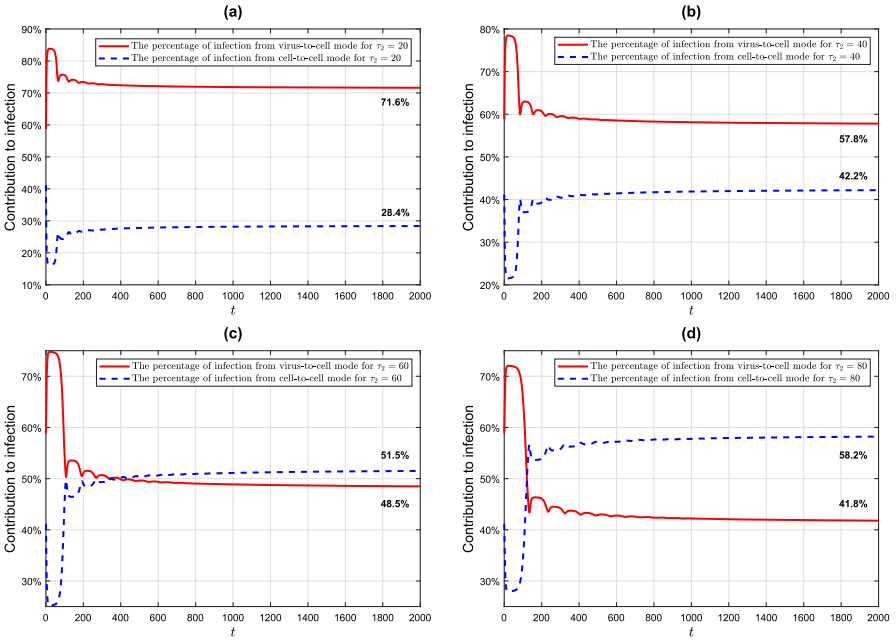

Fig. 3 Contributions of the virus-to-cell and cell-to-cell transmission modes to the total newly produced infected cells with different values of the average delay of virion maturation τ_2

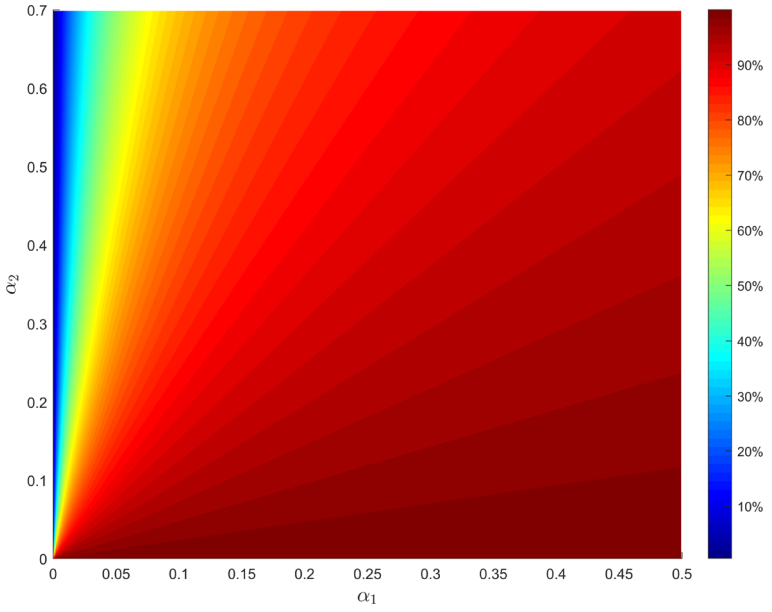

Fig. 4 Contribution of the virus-to-cell transmission mode to the total newly produced infected cells with varying α_1 and α_2. The weak kernels with $\gamma_1 = 1$, $\gamma_2 = 0.5$ and $\gamma_3 = 0.2$ are used

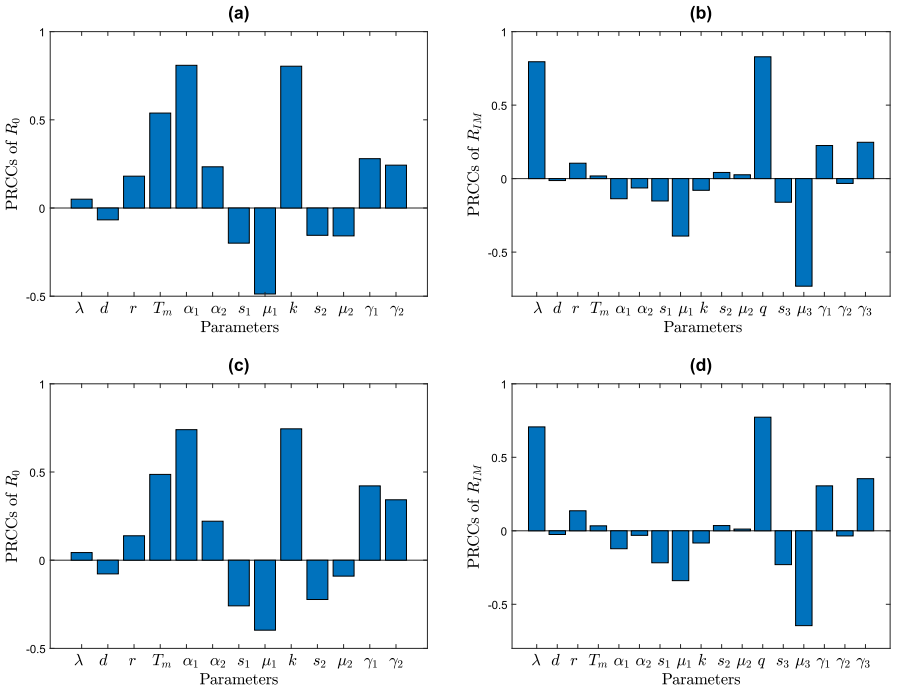

Fig. 5 Sensitivity analysis for the infection reproduction number R_0 and the basic reproduction number for immune response R_{IM} of system (4.1). **a**, **b** The PRCCs for R_0, R_{IM} with weak kernels $\widetilde{f}_i(\tau) = \gamma_i e^{-\gamma_i \tau}$ for $i = 1, 2, 3$. **c**, **d** The PRCCs for R_0, R_{IM} with strong kernels $\widetilde{f}_i(\tau) = \gamma_i^2 \tau e^{-\gamma_i \tau}$ for $i = 1, 2, 3$

see Fig. 5(a)–(c), while the parameters λ and q have strong and positive impact on the basic reproduction number for immune response R_{IM}; see Fig. 5(b)–(d). We also note that R_0 is positively correlated with α_2 and R_{IM} is negatively correlated with α_2, which suggest that ignoring the cell-to-cell transmission mode may underestimate the viral infection level and overestimate the effectiveness of immune response.

4.4 Effects of antiretroviral therapy by a Filippov control

Piecewise smooth differential systems have widely used to construct mathematical models in ecology and medicine (Tang et al. 2012; Wang and Xiao 2013). For example, Wang and Xiao (2013) proposed an epidemic Filippov model, where the treatment strategies were implemented once the susceptible population exceeds a critical value. In our model with antiretroviral therapy, we will consider the treatments of HIV-1 with two major antiretroviral drugs, the reverse transcriptase inhibitors (RTIs) and the protease inhibitors (PIs) (Perelson et al. 1997; Rong and Perelson 2009). RTIs can block the reverse transcription and suppress the process of viral DNA to HIV RNA. PIs can interfere with the replication of HIV protease and induce infected cells to generate noninfectious viral particles. To include the reverse transcriptase inhibitors in our model, we replace α_i $(i = 1, 2)$ in system (4.1) with $\alpha_i(1 - \rho_{RT})$, where $\rho_{RT} \in (0, 1)$ denotes the efficacy of RTIs. Similarly, the viral production rate k is

changed to $k(1-\rho_{PI})$ to incorporate the effects of protease inhibitors, where ρ_{PI} is the efficacy of PIs with $0 < \rho_{PI} < 1$. Following the ideas in (Kuznetsov et al. 2003), we implement the antiretroviral treatment when the viral load V exceeds a threshold value V_c. We also choose weak kernels in (4.1); namely, $\widetilde{f}_i(\tau) = \gamma_i e^{-\gamma_i \tau}$ for $i = 1, 2, 3$. Hence, we reach at the following Filippov system

$$\begin{aligned}
T'(t) &= \lambda - dT(t) + rT(t)\left(1 - \frac{T(t)}{T_m}\right) - \alpha_1(1-\sigma\rho_{RT})T(t)V(t) \\
&\quad - \alpha_2(1-\sigma\rho_{RT})T(t)I(t), \\
I'(t) &= \alpha_1(1-\sigma\rho_{RT})\gamma_1 \int_0^\infty e^{-(s_1+\gamma_1)\tau} T(t-\tau)V(t-\tau)d\tau \\
&\quad + \alpha_2(1-\sigma\rho_{RT})\gamma_1 \int_0^\infty e^{-(s_1+\gamma_1)\tau} T(t-\tau)I(t-\tau)d\tau - \mu_1 I(t) - pI(t)Z(t), \\
V'(t) &= k(1-\sigma\rho_{PI})\gamma_2 \int_0^\infty e^{-(s_2+\gamma_2)\tau} I(t-\tau)d\tau - \mu_2 V(t), \\
Z'(t) &= q\gamma_3 \int_0^\infty e^{-(s_3+\gamma_3)\tau} I(t-\tau)Z(t-\tau)d\tau - \mu_3 Z(t),
\end{aligned} \tag{4.3}$$

where

$$\sigma = \begin{cases} 0, & \text{if } V(t) < V_c, \\ 1, & \text{if } V(t) > V_c. \end{cases} \tag{4.4}$$

We present the results of numerical simulations for the Filippov system (4.3) in Figs. 6, 7, and 8. We observe that enhancing the drug efficacy is effective in lowering the viral loads; see Fig. 6(a). The fluctuating viral loads (with sustained oscillations) can be reduced to a relatively low level with no fluctuations; see Fig. 6(b). In Fig. 7, we observe that the amplitude and the average of the viral load under the uninterrupted therapy (i.e., $V_c = 0$) could be larger than those under the Filippov control with a small but positive V_c. This implies that the therapy strategy with a Filippov control is more efficient than the uninterrupted therapy strategy in reducing the viral loads.

Now, we choose V_c as the bifurcation parameter to investigate the effects of this threshold value on the dynamics of (4.3). It is observed from Fig. 8 that, as V_c increases, the viral load changes from sustained oscillations to a steady state and then back to sustained oscillations. This indicates that the choice of the control threshold V_c is critical to maintain a low level viral load with small fluctuations.

5 Conclusion and discussion

In this paper, we proposed a general delayed viral infection model with both virus-to-cell and cell-to-cell transmission modes. The CTL immune response is also considered. Our model incorporates three distributed time delays for viral infection, viral production and CTLs recruitment, respectively. We have shown that the model admits three possible equilibria: the infection-free equilibrium E_0, the immune-inactivated equilibrium E_1 and the immune-activated equilibrium E_2. Based on the biologically relevant assumptions ($\mathbf{H_1}$)–($\mathbf{H_3}$), we proved that the existence and uniqueness, as well as the

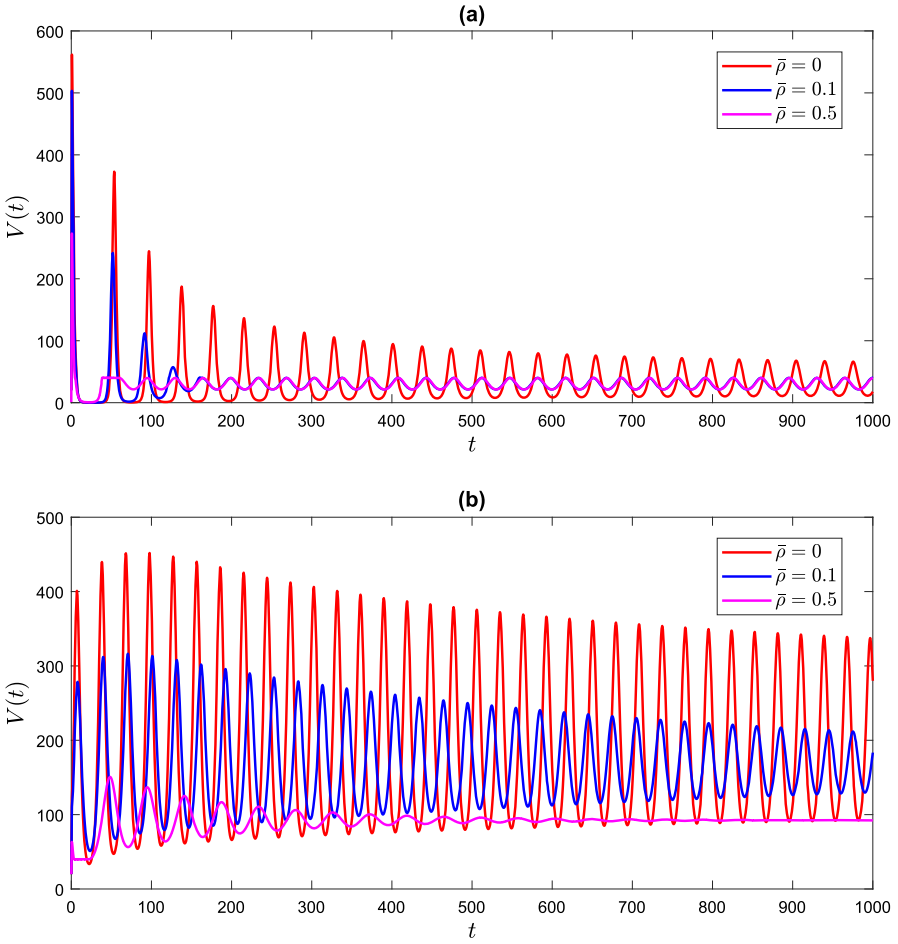

Fig. 6 Effects of drug efficacy on viral loads under a Filippov control. The parameter values are chosen as $V_c = 40$, $\gamma_1 = 1$, $\gamma_2 = 0.5$, and $\rho_{RT} = \rho_{PI} = \bar{\rho}$ varies from 0 to 0.5. The difference between two subfigures lies in the choice of γ_3. In **a** we set $\gamma_3 = 0.1$ while in **b** we choose $\gamma_3 = 0.0125$

global stability of such three equilibria, are determined by two threshold values: the basic reproduction number for infection R_0 and the basic reproduction number for immune response R_{IM}. We have demonstrated that: (1) the viral particles can be cleared out (i.e., E_0 is globally asymptotically stable) if $R_0 \leq 1$; (2) the infection persists and the immune response is absent (i.e., E_1 is globally asymptotically stable) if $R_{IM} \leq 1 < R_0$; and (3) both the viral particles and the immune cells persist (i.e., E_2 is globally asymptotically stable) if $R_{IM} > 1$. Since the basic reproduction number R_0 couples the contributions from both virus-to-cell and cell-to-cell transmissions, it is indispensable to weigh both infection modes *in vivo* to explore the viral dynamics and provide effective containment measures.

The works in (Chen et al. 2016; Shu et al. 2018) have revealed that the intracellular delay and the viral production delay do not generate sustained oscillations under

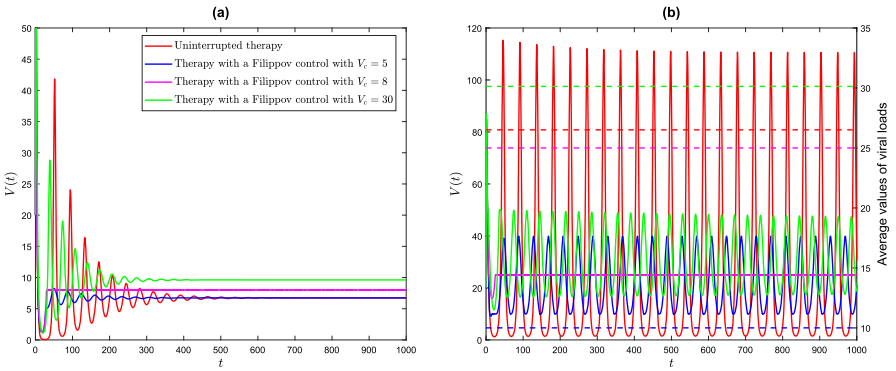

Fig. 7 Comparisons of viral loads under an uninterrupted therapy (i.e., $V_c = 0$) and the Filippov control therapies with different positive threshold values V_c. The parameter values are chosen as $\rho_{RT} = \rho_{PI} = 0.5$, $\gamma_1 = 1$, and $\gamma_2 = 0.5$. In **a** we set $\gamma_3 = 1$ and there is no oscillation, while in **b** we choose $\gamma_3 = 0.1$ and there are sustained oscillations. In (**b**), the red, blue, magenta and green colors correspond to $V_c = 0, 10, 25, 50$, respectively, and the dashed curves denote the corresponding average values of the viral loads (color figure online)

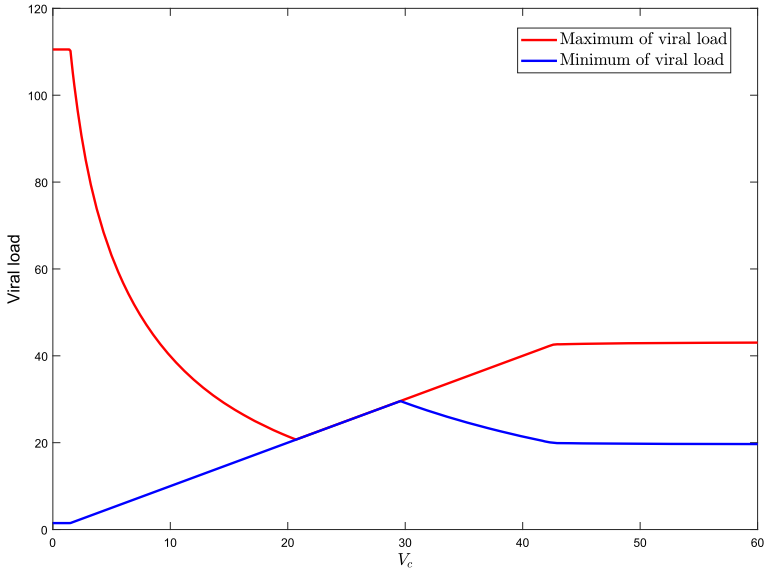

Fig. 8 Bifurcation diagram of (4.3). Parameter values are: $\rho_{RT} = \rho_{PI} = 0.5$, $\gamma_1 = 1$, $\gamma_2 = 0.5$, and $\gamma_3 = 0.1$

certain conditions. Similar results for the proposed model were obtained in Theorems 3.1 and 3.3. We have shown that the value of R_0 decreases as the intracellular delay or the viral production delay increases, which indicates that prolonging the time lags during the processes of viral infection and viral production is beneficial to viral load reduction and viral eradication. For illustration, we selected the kernel functions $\widetilde{f}_i(\tau)$ $(i = 1, 2, 3)$ as the delta function or the gamma distributions, and observed that the

CTLs recruitment delay may induce a bubble bifurcation diagram; namely, as the delay increases, there is a switch from a stable steady state to sustained oscillations, and another switch back to a stable steady state; see Fig. 2.

We also compared the contributions of the virus-to-cell and the cell-to-cell infection modes by calculating the percentages of cumulative infections caused by these two modes. We found that the growth rate of target cells, the average intracellular delay, and the average CTLs recruitment delay have limited impacts on the percentages of infections, but the average viral production delay has a big impact on the relative contributions. In particular, when the average viral production delay is large, the contribution from the cell-to-cell mode may even outweigh that from the virus-to-cell mode. This indicates that the cell-to-cell transmission route may play a significant role in the viral dynamics and hence, we may need to find a new direction for antiviral treatment that blocks the cell-to-cell infections. We also showed that antiretroviral treatment strategies with a Filippov control can be used to maintain the viral load at a low level by choosing a suitable control threshold value.

Acknowledgements We would like to thank two anonymous referees and the editor for their careful reading and helpful suggestions which led to an improvement to our original manuscript. H. Shu is partially supported by the National Natural Science Foundation of China (No. 11971285), and the Fundamental Research Funds for the Central Universities (No. GK202201002). L. Wang is partially supported by the Natural Sciences and Engineering Research Council of Canada (No. RGPIN-2020-04143). X.-S. Wang is partially supported by the Louisiana Board of Regents Support Fund under contract No. LEQSF(2022–25)-RD-A-26.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Bonhoeffer S, May RM, Shaw GM, Nowak MA (1997) Virus dynamics and drug therapy. Proc Natl Acad Sci USA 94:6971–6976

Canabarro AA, Gléria IM, Lyra ML (2004) Periodic solutions and chaos in a non-linear model for the delayed cellular immune response. Physica A 342:234–241

Chen SS, Cheng CY, Takeuchi Y (2016) Stability analysis in delayed within-host viral dynamics with both viral and cellular infections. J Math Anal Appl 442:642–672

Dahari H, Lo A, Ribeiro RM, Perelson AS (2007) Modeling hepatitis C virus dynamics: liver regeneration and critical drug efficacy. J Theor Biol 247:371–381

Dixit NM, Perelson AS (2004) Multiplicity of human immunodeficiency virus infections in lymphoid tissue. J Virol 78:8942–8945

Galloway NLK, Doitsh G, Monroe KM, Yang Z, Muñoz-Arias I, Levy DN, Greene WC (2015) Cell-to-cell transmission of HIV-1 is required to trigger pyroptotic death of lymphoid-tissue-derived CD4 T cells. Cell Rep 12:1555–1563

Gummuluru S, Kinsey CM, Emerman M (2000) An in vitro rapid-turnover assay for human immunodeficiency virus type 1 replication selects for cell-to-cell spread of virus. J Virol 74:10882–10891

Guo T, Qiu Z, Rong L (2020) Analysis of an HIV model with immune responses and cell-to-cell transmission. Bull Malays Math Sci Soc 43:581–607

Hale JK, Kato J (1978) Phase space for retarded equations with infinite delay. Funkc Ekvac 21:11–41

Hale JK, Verduyn Lunel SM (1993) Introduction to functional differential equations. Springer-Verlag, New York
Hale JK, Waltman P (1989) Persistence in infinite-dimensional systems. SIAM J Math Anal 20:388–395
Herz AV, Bonhoeffer S, Anderson RM, May RM, Nowak MA (1996) Viral dynamics in vivo: limitations on estimates of intracellular delay and virus decay. Proc Natl Acad Sci USA 93:7247–7251
Huang G, Takeuchi Y, Ma W (2010) Lyapunov functional for delay differential equations model of viral infections. SIAM J Appl Math 70:2693–2708
Hübner W, McNerney GP, Chen P, Dale BM, Gordan RE, Chuang FYS, Li XD, Asmuth DM, Huser T, Chen BK (2009) Quantitative 3D video microscopy of HIV transfer across T cell virological synapses. Science 323:1743–1747
Iwami S, Takeuchi JS, Nakaoka S, Mammano F, Clavel F, Inaba H, Kobayashi T, Misawa N, Aihara K, Koyanagi Y, Sato K (2015) Cell-to-cell infection by HIV contributes over half of virus infection. eLife 4:e08150
Komarova NL, Anghelina D, Voznesensky I, Trinite B, Levy DN, Wodarz D (2013) Relative contribution of free-virus and synaptic transmission to the spread of HIV-1 through target-cell populations. Biol Lett 9:1049–1055
Komarova NL, Levy DN, Wodarz D (2013) Synaptic transmission and the susceptibility of HIV infection to anti-viral drugs. Sci Rep 3:1–8
Kuznetsov YA, Rinaldi S, Gragnani A (2003) One-parameter bifurcations in planar Filippov systems. Int J Bifurc Chaos 13:2157–2188
Lai X, Zou X (2014) Modeling HIV-1 virus dynamics with both virus-to-cell infection and cell-to-cell transmission. SIAM J Appl Math 74:898–917
Leenheer PD, Smith HL (2003) Virus dynamics: a global analysis. SIAM J Appl Math 63:1313–1327
Li MY, Shu H (2010) Impact of intracellular delays and target-cell dynamics on in vivo viral infections. SIAM J Appl Math 70:2434–2448
Liu M, Liz E, Röst G (2015) Endemic bubbles generated by delayed behavioral response: global stability and bifurcation switches in an SIS model. SIAM J Appl Math 75:75–91
Martin N, Sattentau Q (2009) Cell-to-cell HIV-1 spread and its implications for immune evasion. Curr Opin HIV AIDS 4:143–149
Nakata Y (2010) Global dynamics of cell mediated immunity in viral infection models with distributed delays. J Math Anal Appl 375:14–27
Neumann AU, Lam NP, Dahari H, Gretch DR, Wiley TE, Layden TJ, Perelson AS (1998) Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. Science 282:103–107
Nowak MA, Bangham CRM (1996) Population dynamics of immune responses to persistent viruses. Science 272:74–79
Perelson AS, Essunger P, Cao Y, Vesanen M, Hurley A, Saksela K, Markowitz M, Ho DD (1997) Decay characteristics of HIV-1-infected compartments during combination therapy. Nature 387:188–191
Perelson AS, Kirschner DE, De Boer R (1993) Dynamics of HIV infection of CD4+ T cells. Math Biosci 114:81–125
Perelson AS, Nelson PW (1999) Mathematical analysis of HIV-1 dynamics in vivo. SIAM Rev 41:3–44
Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582–1586
Ribeiro RM, Lo A, Perelson AS (2002) Dynamics of hepatitis B virus infection. Microbes Infect 4:829–835
Rong L, Perelson AS (2009) Modeling HIV persistence, the latent reservoir, and viral blips. J Theor Biol 260:308–331
Sato H, Orenstein J, Dimitrov D, Martin M (1992) Cell-to-cell spread of HIV-1 occurs within minutes and may not involve the participation of virus particles. Virology 186:712–724
Shu H, Chen Y, Wang L (2018) Impacts of the cell-free and cell-to-cell infection modes on viral dynamics. J Dyn Differ Equ 30:1817–1836
Shu H, Wang L, Watmough J (2013) Global stability of a nonlinear viral infection model with infinitely distributed intracellular delays and CTL immune responses. SIAM J Appl Math 73:1280–1302
Shu H, Xu W, Wang X-S, Wu J (2020) Complex dynamics in a delay differential equation with two delays in tick growth with diapause. J Differ Equ 269:10937–10963
Sigal A, Kim JT, Balazs AB, Dekel E, Mayo A, Milo R, Baltimore D (2011) Cell-to-cell spread of HIV permits ongoing replication despite antiretroviral therapy. Nature 477:95–98
Sourisseau M, Sol-Foulon N, Porrot F, Blanchet F, Schwartz O (2007) Inefficient human immunodeficiency virus replication in mobile lymphocytes. J Virol 81:1000–1012

Tang S, Liang J, Xiao Y, Cheke RA (2012) Sliding bifurcations of Filippov two stage pest control models with economic thresholds. SIAM J Appl Math 72:1061–1080
Thieme HR (2009) Spectral bound and reproduction number for infinite-dimensional population structure and time heterogeneity. SIAM J Appl Math 70:188–211
Tsiang M, Rooney JF, Toole JJ, Gibbs CS (1999) Biphasic clearance kinetics of hepatitis B virus from patients during adefovir dipivoxil therapy. Hepatology 29:1863–1869
Wang A, Xiao Y (2013) Sliding bifurcation and global dynamics of a Filippov epidemic model with vaccination. Int J Bifurc Chaos 23:1350144
Wang J, Guo M, Liu X, Zhao Z (2016) Threshold dynamics of HIV-1 virus model with cell-to-cell transmission, cell-mediated immune responses and distributed delay. Appl Math Comput 291:149–161
Wang K, Wang W, Pang H, Liu X (2007) Complex dynamic behavior in a viral model with delayed immune response. Physica D 226:197–208
Wang Y, Zou Y, Brauer F, Heffernan J (2013) Viral dynamics model with CTL immune response incorporating antiretroviral therapy. J Math Biol 67:901–934
Yang Y, Zhang T, Xu Y, Zhou J (2017) A delayed virus infection model with cell-to-cell transmission and CTL immune response. Int J Bifurc Chaos 27:1750150